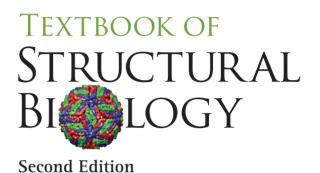
TEXTBOOK OF

STRUCTURAL BIOLOGY

Second Edition

Anders Liljas
Lars Liljas
Miriam-Rose Ash
Göran Lindblom
Poul Nissen
Morten Kjeldgaard





Series in Structural Biology

Series Editor: Anders Liljas (Lund University, Sweden)

| ъ. | | | |
|----|---|----|-----|
| Pn | h | 18 | hed |

- Vol. 1 Structural Aspects of Protein Synthesis (Second Edition) by Anders Liljas and Måns Ehrenberg
- Vol. 2 Found in Translation: Collection of Original Articles on Single-Particle Reconstruction and the Structural Basis of Protein Synthesis by Joachim Frank
- Vol. 3 Selected Papers of Michael G. Rossmann with Commentaries: The Development of Structural Biology by Michael G. Rossmann
- Vol. 4 From a Grain of Salt to the Ribosome: The History of Crystallography as seen Through the Lens of the Nobel Prize edited by Ivar Olovsson, Anders Liljas and Sven Lidin
- Vol. 5 The Struggles and Dreams of Robert Langer by Robert Langer
- Vol. 6 Structure and Action of Molecular Chaperones:
 Machines that Assist Protein Folding in the Cell
 by Lila M. Gierasch, Arthur L. Horwich, Christine Slingsby, Sue Wickner and
 David Agard
- Vol. 7 Curiosity and Passion for Science and Art by Uwe B. Sleytr
- Vol. 8 Textbook of Structural Biology (Second Edition)
 by Anders Liljas, Lars Liljas, Miriam-Rose Ash, Göran Lindblom, Poul Nissen and
 Morten Kjeldgaard

TEXTBOOK OF STRUCTURAL BIOLOGY

Second Edition

Anders Liljas (Lund University, Sweden)

Lars Liljas (Uppsala University, Sweden)

Miriam-Rose Ash (The University of Sydney, Australia)

Göran Lindblom (Umeå University, Sweden)

Poul Nissen (Aarhus University, Denmark)

Morten Kjeldgaard (Aarhus University, Denmark)

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601 UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Names: Liljas, Anders, author.

Title: Textbook of structural biology / Anders Liljas, Lund University, Sweden [and five others].

Description: Second edition. | [Hackensack] New Jersey: World Scientific, 2017. |

Series: Series in structural biology; vol. 8 | Includes bibliographical references and index.

Identifiers: LCCN 2016033861 | ISBN 9789813142466 (hardcover : alk. paper) |

ISBN 9813142464 (hardcover : alk. paper) | ISBN 9789813142473 (pbk. : alk. paper) |

ISBN 9813142472 (pbk. : alk. paper)

Subjects: LCSH: Proteins--Structure. | Ultrastructure (Biology) | Protein folding. | Biomolecules--Structure.

Classification: LCC QP551 .T45 2017 | DDC 572/.633--dc23 LC record available at https://lccn.loc.gov/2016033861

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Copyright © 2017 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

We dedicate this textbook to the memory of our coauthor Jure Piskur.



Contents

| Preface to the | e Second Edition | ix |
|----------------|---|-----|
| Chapter 1 | Introduction | 1 |
| Chapter 2 | Basics of Protein Structure | 11 |
| Chapter 3 | The Folding, Folds and Functions of Proteins | 37 |
| Chapter 4 | Basics of Membrane Proteins | 69 |
| Chapter 5 | Basics of Nucleic Acid Structure | 105 |
| Chapter 6 | Basics of Lipids and Membrane Structure | 161 |
| Chapter 7 | Basics of Carbohydrates | 213 |
| Chapter 8 | Enzymes | 227 |
| Chapter 9 | Genome Structure, DNA Replication and Recombination | 269 |
| Chapter 10 | Transcription | 307 |
| Chapter 11 | Protein Synthesis — Translation | 351 |
| Chapter 12 | Protein Folding and Degradation | 385 |
| Chapter 13 | Transmembrane Transport | 425 |
| Chapter 14 | Signal Transduction | 451 |
| Chapter 15 | Cell Motility and Transport | 481 |
| Chapter 16 | Structural Aspects of Cell-Cell Interactions | 507 |
| Chapter 17 | The Immune System | 521 |

viii ■ Textbook of Structural Biology, 2nd Edition

| Chapter 18 | Virus Structure and Function | 535 |
|------------|--|-----|
| Chapter 19 | Bioinformatics Tools in Structural Biology | 553 |
| Index | | 577 |

Preface to the Second Edition

It is now over seven years ago since we published the first edition of our *Textbook of Structural Biology*. Our coauthor of the first edition, Jure Piskur, died prematurely in 2014 due to cancer. We very much miss his insights and skills.

The field has made very rapid and extensive progress since the first edition was released. In particular, it is becoming evident that many molecular systems are highly integrated with each other and sometimes very large assemblies are formed. This complexity is gradually emerging and accessible for structural investigations. The field of membrane proteins is making remarkable progress and is more extensively covered in this edition with its own dedicated chapter. Furthermore, we have added a chapter on carbohydrates, which is an emerging area that is deeply integrated with protein and cellular structure and function.

Given the impressive number of unique protein structures now available, it is of course impossible to cover all of these in any single volume. We have therefore tried to concentrate on the growing body of lasting knowledge of the structure and the roles of macromolecular systems.

We have received help from many colleagues who have used the first edition of the textbook for courses. In addition, we have had expert help in reviewing certain chapters. We are indebted to Prof Ulf Lindahl for expert insight into carbohydrates. Furthermore, we wish to express our thanks to Lars Erik Andreas Ehnbom and Saraboji Kadhirvel for expert skills in producing parts of the illustrations. Many illustrations are produced using the program Molscript by Per Kraulis.

Introduction

1.1 <u>Life</u>

We are surrounded by microbes, plants and animals that we immediately recognize as living beings (Figure 1.1). However, it is still difficult to provide a concise definition what life is. Perhaps the most useful definition for this book is that life is a unit capable of chemical activities, which can reproduce and evolve.

Chemical activities, which involve conversions of energy and matter, are called *metabolism*. These activities capture energy and chemical matter in different forms. Thousands of chemical activities take place simultaneously in a living organism and they must be well coordinated or regulated to maintain the stability of the living unit.

Reproduction of the unit (generating new units) provides both the continuity and the variation that is also an important characteristic of life. The combination of reproduction, horizontal transfer of information and "erroneous" duplicates provides the basis of evolution. In other words, the composition of the unit should be able to change over time to better adapt to the changing environmental conditions. Living organisms appear in very different forms and follow very different life-styles. However, the basic characteristics of life (including metabolism, reproduction and evolution) are provided and governed by very similar sub-structures: biological macromolecules and cells.

1.2 Levels of Organization of Life

The living world has several hierarchical levels, ordered from the smallest to the largest. At the bottom are molecules, a mix of inorganic and organic compounds and biological



Fig. 1.1 ■ Living organisms are found in numerous different forms. *Left*: A microscope picture of baker's yeast (*Saccharomyces cerevisiae*) cells (by the courtesy of Concetta Compagno). *Right*: Linneas (*Linnea borealis*) covering vast areas of Lapland (by the courtesy of Bernarda Rotar) and *bottom*: moose, the largest land animals in Scandinavia (by the courtesy of Aca.Pixus.dk). Within these different macro-forms very similar molecular structures can be found, which determine the form and lifestyle of the carrier organisms.

macromolecules, followed by sub-cellular structures, cells, tissues, organs, organisms, populations, communities and the biosphere, which encompasses all biological communities on the Earth.

Macromolecules are central in all living organisms. They are giant polymers consisting of repeating units. These repeating units may or may not be identical, and are connected with covalent or non-covalent bonds. Macromolecules perform a multitude of functions, which are the basis of metabolism, reproduction and evolution, such as energy or information storage, reaction catalysis, coordination and regulation, communication, structural

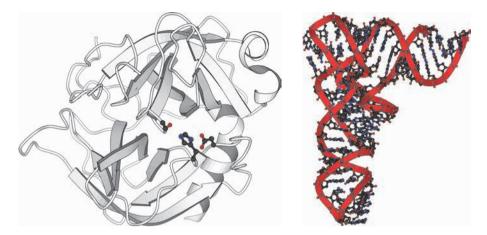


Fig. 1.2 ■ A simplified picture of two bio-macromolecules, which are the focus of our further chapters. *Left:* The structure of a well-known protein, chymotrypsin (PDB: 4CHA). *Right:* A nucleic acid molecule, yeast tRNAPhe (PDB: 1EHZ).

support, defense, movement and transport. On the basis of chemical composition we talk about three different kinds of macromolecules: (i) peptides and proteins that are polymers of amino acid residues, (ii) nucleic acids, which are polymers of nucleotides, and (iii) carbohydrates, which are polymers of sugars (Figure 1.2). Other central molecules that should be mentioned here are the lipids. Although they are not macromolecules, they self-assemble into large aggregates of macromolecular dimensions, including the lipid bilayer (an important building block of cell membranes), micellar aggregates containing bile molecules, and the aggregates of lipoproteins that transport cholesterol and fat in the blood stream. In the following chapters we will try to understand the structures of bio-macromolecules and link them to their functions and the higher levels of the living world.

The basic unit of life is a *cell* (Figure 1.3). Cells are surrounded by a plasma membrane, which separates each cell from the external environment and creates a segregated compartment with a controlled internal environment. Cells show two organizational patterns: (i) prokaryotic, characteristic for Bacteria and Archaea, and (ii) eukaryotic, characteristic for Eukarya. Prokaryotic cells usually exist as single cells and are smaller than eukaryotic ones, typically on the order of 1 µm in diameter. The basic structure of a prokaryotic cell is defined by a cellular membrane, an intracellular nucleoid containing DNA, and the cytosol holding the rest of the intracellular material, where ribosomes, enzymes and cytoskeletal elements are found. Eukaryotic cells are usually at least ten times larger than prokaryotic cells and more complex, with inner membranes separating compartments and organelles. The organelles include: (i) the nucleus, storing genetic material and the replication and gene transcription systems (ii) the cytosol, where protein synthesis and many essential biochemical reactions take place, (iii) the mitochondrion, a power plant and energy storage compartment, (iv) the endoplasmatic reticulum and Golgi apparatus, where proteins are matured and sorted to further locations, (v) the lysosomes or vacuoles, where polymeric macromolecules, such as proteins, are recycled into usable metabolites.

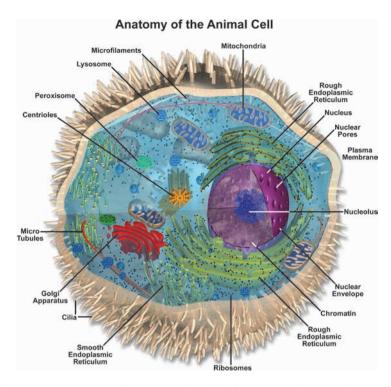


Fig. 1.3 ■ A schematic picture of an animal cell showing sub-cellular structures, such as nucleus, membrane systems (ER), mitochondrion, etc. (Made by Michael W. Davidson, Florida State University.)

All organisms on Earth seem to originate from a single unicellular *organism*. The main reasons why one can claim that all organisms originate from the same cell is that not only do all living species use the same nucleotides and amino acids despite many other possibilities, but the genetic code (the dictionary for translation from the language of nucleic acids to the one of proteins) is the same. In addition, central molecular systems like transcription and translation are strongly related. A smaller molecule like ATP is the universal currency of energy in all living organisms, although in principle many other choices would have been possible.

Today, many millions of different organisms that do not interbreed with each other are found and we call them species (Figure 1.4). They are all adapted to their different environments and in a naive sense they may seem perfect. However, a particular life form may not be fit tomorrow and thereby become extinct, like so many other species in the past, which have previously populated Earth. Due to changes of environment, new and better-fit species constantly evolve over time, and this evolution works by gradually changing the structures of macromolecules.

The unfolding of events leading to the present diversity is expressed as an evolutionary tree showing the order in which species split and evolved into new species. This tree traces the descendants coming from ancestors that lived at different times in the past.

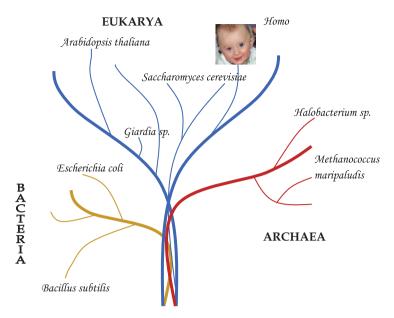


Fig. 1.4 ■ A simplified tree of life. The common progenitor originated approximately 4 billion years ago. The position of the first branchings, occurring between the progenitor of Bacteria, Archaea and Eukarya, are still unclear.

In other words, the evolutionary tree shows the evolutionary relationship among modern and ancient species. It is important to understand the evolutionary relationship between organisms when one compares the structure of macromolecules from these organisms, being involved in similar processes. Some of the earlier branching is difficult to reconstruct because there are no available fossils. However, based on molecular evidence in modern organisms, we can separate all living organisms into three domains, which have been evolving separately for more than 1 billion years: (i) Archaea, (ii) Bacteria, and (iii) Eukarya. Even if they superficially look similar, Archaea and Bacteria separated into distinct lineages very early during evolutionary history.

Archaea are often inhabitants of extreme environments, such as hot and acidic springs, sea depths and salt brines, but can also be found in more "normal" environments. Their replication, transcription and translation machinery resembles the eukaryotic machinery, while their metabolism and energy conversion resemble the bacterial ones.

Bacteria consist of more than a dozen sub-groups, also called clades, but the most important are: Protobacteria, Cyanobacteria, Spirochetes, Chlamydias and Firmicutes. The Protobacteria are the largest and a very diverse group, including one of the best-studied organisms, *Escherichia coli*. Sometimes bacteria are divided, on the basis of their cell wall composition, into gram-positive, including *Bacillus subtilis*, and gram-negative, including *E. coli*. Bacteria exhibit the greatest biochemical diversity.

Eukarya can be divided into four groups: Protista, Plantae, Fungi and Animalia. The Protista contain mostly single celled organisms and have a polyphyletic origin, meaning

that some represent very primitive eukaryotes, such as *Giardia*, while some are closely related to animals, such as *Dictyostelium*, or plants, such as red algae. Phagotrophy, a feeding mode to form a pocket in the plasma membrane and enclose the "food", is a hallmark of Eukarya.

1.3 Short History of Life on Earth

The theory of chemical evolution holds that conditions on the primitive Earth, around 4 billion years ago, led to the emergence of the first biological molecules. Oparin and Haldane independently suggested in the 1920s that if Earth's first atmosphere was reducing, and if there was a supply of external energy then a range of organic compounds might be synthesized. In the 1950s, Stanley Miller and Harold Urey mimicked these conditions in the lab. Water vapor, hydrogen gas, ammonia and methane gas were exposed to sparks, and after a few days the system contained several complex molecules, such as amino acids and nucleic acid bases, the building blocks of today's life. When the monomeric units were present, it was not so difficult to achieve polymerization even under abiotic conditions. However, how could the first peptides and nucleic acids become "alive"? In other words, how could they start reproducing and evolving?

The term *replicator* means a structure that can arise only if there is a preexisting structure of the same kind in the vicinity. For example, a supersaturated solution crystallizes if a small seeding crystal is added. However, this represents a simple replicator relying on a single structure. More sophisticated replicators could exist in several forms and thereby could have contributed to heredity. For sustained evolution, an indefinite number of forms and indefinite variation in heredity is necessary. The first artificial replicator, a simple hexadeoxynucleotide not needing enzymes for its replication (polymerization from the present mono-units) was synthesized by von Kiedrowski in 1986. The first short RNA molecules may have had the ability to catalyze the polymerization of offspring molecules. The first replicating RNA molecules competed successfully with their own erroneous copies and with other less-efficient systems for the monomers needed for their replication. Even if self-replicating RNA molecules fulfill the above criteria for life, the path to the first cells was still more sophisticated. One of the main following steps was to include peptides and proteins to establish the RNA — protein world, followed by the introduction of membrane systems, thereby segregating the primitive cell from its environment.

The origin of the first cell, the common progenitor of all living organisms, could be approximately 3.5 billion years ago, and at that time simple replication and translation machineries already existed. One hypothesis suggests that, during the following 2 billion years, the unicellular system evolved to represent a fine net of metabolic reactions connected to increasingly sophisticated machineries for nucleic acid replication and RNA to protein translation, also keeping plasticity, enabling the cell to respond to the demands of

the ever-changing environment. During this period, the first living cells were still dependent on organic compounds, which were the primary source of energy, and had abiotic origin. Later, approximately 2.5 billion years ago, one of the major steps was the evolution of the ability to use the energy of sunlight to power metabolism. Photosynthesis provided energetic independence and soon resulted in vast quantities of organic materials and oxygen. The evolution of aerobic metabolism significantly changed cellular biochemistry. Many enzymatic reactions became dependent, directly or indirectly, on the presence of oxygen. Aerobic metabolism allowed cells to grow larger. Some of the further major transitions include the origin of sex, the origin of multicellular organisms and the origin of social groups. Behind all these events stood proteins, nucleic acids, carbohydrates and lipids, with their evolving structures and functions.

1.4 What is Structural Biology and When Did It Start?

The field of structural biology focuses on a classical insight: in order to understand, we need to see. "Seeing is believing", or "a picture says more than a thousand words" are well-known phrases. This is true whether we deal with large objects, as in astronomy and astrophysics, medium-sized objects such as birds or fishes, or with very small objects like biochemical systems or particle physics. Structural biology is the science that tries to make the sub-cellular and molecular objects of biology visible and understandable.

It is difficult to identify the very beginning of structural biology. One important step is the purification of the fundamental components. Friedrich Miescher discovered and isolated DNA in 1869. The understanding of the biological role of DNA did not start until 1944 when Avery, MacLeod and McCarty showed that DNA is the genetic material. Elucidation of the structure of DNA in 1953 was a major milestone in structural biology. Francis Crick and James Watson, using diffraction data obtained by Rosalind Franklin and Maurice Wilkins, deduced a model for DNA. This model led to a detailed insight of the replication of DNA, the transcription of DNA to RNA and also the key steps of translation, central activities in molecular biology.

In a review from 1964, James Watson expressed: "Unfortunately, we cannot accurately describe at the chemical level how a molecule functions unless we know first its structure." This describes the situation in a nutshell and it is exemplified in one field of biology after the other.

Proteins have long been known, but the molecular nature of them was poorly understood. Jöns Jacob Berzelius (1779–1848), the well-known Swedish chemist, introduced the term protein. Proteins were classified as colloids without defined structures and shapes. The first crystallization of proteins may have been of hemoglobin, in 1840 by Hünefeld.

At this time, the crystals were called "blood crystals" and it was not realized that the red crystals were built of a protein. Several other proteins were also crystallized early on. The nature of proteins became better understood when Theodor (The) Svedberg could show with his ultracentrifuge that proteins have unique molecular weights.

During the 19th century the action of gastric juice on the degradation of solid proteins was thoroughly studied. One active ingredient was called pepsin but its nature was entirely unclear. Gradually, the catalytic substance was given names as "ferment" or "enzyme". Enzymes were believed by (among others) the Nobel laureate Willstätter to be of a different nature than lipids, carbohydrates or proteins and present in only very low concentrations in plants or animals. J.B. Sumner and subsequently J.H. Northrop showed that enzymes are proteins with unique structures, since urease and pepsin could be purified and crystallized. John D. Bernal and Dorothy Crowfoot (later Hodgkin) could show that pepsin crystals diffracted X-rays when kept in a moist environment, thus demonstrating that proteins would have a specific structure, which was lost if dried out. In the same period, F.C. Bawden, N.W. Pirie and W.M. Stanley crystallized a number of viruses. The perfection of crystallization and the crystallographic analysis of protein and enzyme crystals took several decades until it matured in the well resolved crystallographic structures of myoglobin and hemoglobin, in 1959 and 1968, respectively.

Structural biology includes a number of methods in addition to diffraction and scattering methods. In an early phase, electron microscopy was already an important technique to obtain an insight into the organization of biological systems and macromolecules. One major advance was the analysis of virus particles by Caspar and Klug in the end of the 1950s and beginning of the 1960s. The symmetry principles could be deduced, and for the larger viruses different functional components were identified. Another development in the field of electron microscopy was the studies of 2D crystals of bacteriorhodopsin studied by R. Henderson and N. Unwin using electron diffraction. This opened new possibilities, but only a limited range of objects yielded material good enough for structural studies. Subsequent to these developments, the single particle reconstruction studies of large molecular complexes at cryo-temperatures (cryo-EM) and tomography have very significantly extended the capabilities of electron microscopy to contribute entirely new insights of structural biology at a range of resolutions. The single particle reconstruction has become a new revolution in the field, with a capacity to analyze complexes that were poorly or not at all available before at high resolution. In addition, cryo-EM has the capacity to identify several different conformations in a single sample, adding insights into the dynamics of functional molecules.

Structural biology has moved numerous systems from an understanding where the molecules are represented by blobs to where the atomic coordinates are available, as well as details of molecular interactions.

A limitation of crystallography is that it gives still pictures of the molecular systems studied, with only limited insight into dynamics. In fortunate cases, a number of states can be crystallized and characterized at atomic resolution. However, in many cases one

would need insights into states that are not accessible to crystallization, maybe because they are too short-lived, to understand the dynamics of the system. Here, NMR-studies can sometimes give the information that is missing.

Generally, NMR spectroscopy can provide structures, which are particularly valuable when crystals cannot be obtained. When both NMR and crystallographic structures are available, the quality of the crystallographic information is normally better. However, the unique contributions of NMR spectroscopy come from dynamic studies of systems where the structures are already known. Here, the mobility and details of transient interactions can be characterized.

To obtain optimal information several methods should be employed. Wrong or partial information can be corrected or extended. In the best-understood systems, physical and theoretical chemists have contributed their experimental or computational methods to get additional angles on the understanding of the system.

A Short Summary of the Book

This book is a textbook of structural biology for undergraduate and graduate students. The focus is to cover the central and most interesting aspects of structure, combined with a focus on interesting biology. The book makes no attempt to cover the entire fields of biology or molecular biology. One selection principle is a focus on systems where we know the structures reasonably well.

We try to provide a comprehensive coverage of the structural and functional understanding of proteins, nucleic acids, lipids and membranes as well as carbohydrates. The book makes no attempt at describing the methods used to obtain the results. This would require a separate volume. In addition to the basic structural knowledge of protein, nucleic acid, lipids and carbohydrates there is a significant coverage of the steps involved in the expression of the genetic information in DNA into proteins. Likewise, the breakdown of macromolecules is covered. Much of biology is related to membranes. They enclose cells or cellular compartments. The passage of material and information across membranes is crucial for all cellular biology and structural insights are rapidly increasing. In multicellular organisms, cell-cell contacts and interactions are essential for coordinated activities. In relation to this, the insights into how cells and organisms move is increasingly better understood. These fields are described in the text.

Insights into the evolution of biological systems and functional genomics also benefit from structural studies. The rate by which the DNA sequences of complete genomes are produced has generated an enormous database that is rapidly growing. DNA or protein sequences can confidently be identified as long as the sequence identity is reasonably good. In many cases, the sequences themselves are insufficient. However, the structural relationship can both suggest the evolutionary relationship of a protein as well as its function. The final chapter gives some approaches to use various predictive methods to access structural insights from sequence data alone.

Further Reading

Avery GT, Macleod CM, McCarty M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonuleic acid fraction isolated from pneumococcus type iii. *J Exp Med* **79**: 137–158.

Bernal JD, Crowffot DC. (1934) X-ray photographs of crystalline pepsin. *Nature* **133**: 794–795.

Franklin RE, Gosling RG. (1953) Molecular configuration in sodium thymonucleate. *Nature* **171**: 740–741.

Fruton JS. (2002) A history of pepsin and related enzymes. Quart Rev Biol 77: 127–147.

Kendrew JC, Dickerson RE, Strandberg BE, *et al.* (1960) Structure of myoglobin: a thre-dimensional Fourier synthesis at 2 A resolution. *Nature* **185**: 442–427.

McPherson A. (1991) A brief story of crystal growth. J Cryst Growth 110: 1–10.

Miescher F. (1871) Über die chemische Zusammensetzung der Eiterzellen. Hoppe-Seyler's medicinisch-chemische Untersuchungen 4: 441–460.

Northrop JH. (1929) Crystalline pepsin, Science 69: 580.

Olofsson I, Liljas A, Lidin S. (2014) From a grain of salt to the ribosome. The history of crystallography as seen through the lens of the Nobel Prize. World Scientific, Singapore.

Sumner JB. (1926) The isolation and crystallization of the enzyme urease: preliminary paper. *J Biol Chem* **69**: 435–441.

Watson JD, Crick FH. (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.

Basics of Protein Structure

2.1 Bonds and Interactions

In most biological processes on the molecular, cellular or organism level, proteins are the main actors, and it is therefore important to understand their function. The function of a protein is determined by its structure. Most proteins are designed to bind other proteins, DNA, RNA or other molecules, and this requires that they form and maintain an exact spatial organization of functional groups. Even though most proteins have a well-defined conformation they are not rigid. Various types of conformational flexibility (side chain and loop movements, domain rotations) are often crucial for their function.

Proteins are extremely plastic. They can adopt a very large variety of shapes. Most positions of a protein sequence can be changed while maintaining the structure and the function. Proteins can be globular or fibrous, stiff or elastic. They can change their conformation in subtle ways or extensively. They can be enzymes, synthesizing small or large molecules but also catalyzing their breakdown. They can be motors, generating rotational or sliding motions. In this and subsequent chapters we will describe the characteristics of proteins that can lead to such a multitude of roles.

Before going into the details of the structure of proteins it may be valuable to review some relevant principles of chemistry that are involved.

2.1.1 Covalent Bonds

A chemical bond between two atoms forms if the resulting arrangement of the two nuclei and their electrons has a lower energy than the total energy of the separate atoms. If the lowest energy is obtained by a complete transfer of one or more electrons from one atom to the other, ions form and the compound is held together by electrostatic

TABLE 2.1 Some Average Bond Lengths

| Bond | Average Length (Å) | | | |
|---------------|--------------------|--|--|--|
| С-Н | 1.09 | | | |
| C-C | 1.54 | | | |
| C-C (benzene) | 1.39 | | | |
| C=C | 1.34 | | | |
| C=C | 1.20 | | | |
| C-O | 1.43 | | | |
| C=O | 1.12 | | | |
| О-Н | 0.96 | | | |
| N-H | 1.01 | | | |
| N-O | 1.40 | | | |
| N=O | 1.20 | | | |

(coulombic) attraction between the ions. Here we have got an ionic bond, e.g. as for NaCl salt crystals. If the lowest energy can be achieved by sharing electrons, then the atoms link through a covalent bond and discrete molecules are formed, such as H₂ and NH₃. For biological macromolecules the covalent bond is the most frequently occurring one. Hydrogen bonds are also very prevalent (see Section 2.1.4) and ionic bonds are sometimes found in so-called salt bridges on the surface of many proteins. When a covalent bond forms, the atoms share electrons until they reach a so-called noble-gas configuration (the octet rule according to Lewis). The changes in energy responsible for the formation of bonds occur when the electrons in the outermost shells (the valence electrons) move to new locations, i.e. the electronic structures of the atoms play a crucial role in bond formations. For a theoretical description of chemical bonds one has to use molecular quantum mechanics. Some experimentally observed typical bond lengths are given in Table 2.1.

Concerning the strength of a bond one can look at the average bond energies, sometimes also called the bond dissociation energy, which is the energy required to break one mole of the particular bond under discussion. Bond energies vary between about 160 to 1100 kJ/mol, depending on the number of double bonds between the atoms (N_2 has a triple bond with 942 kJ/mol), or if there is a large part of ionic charge (as in F_2 with 155 kJ/mol). Certain systematic trends with changes in atomic number are evident. Bonds generally get weaker with increasing atomic number like in the series HF>HCl> HBr>HI. Bond energies, like bond lengths, are fairly reproducible (within about 10%) from one compound to another. It is therefore possible to tabulate average bond energies from measurements on a series of compounds. As a final example, it could be mentioned that the average bond energy increases from 345 to 809 kJ/mol from a single carbon-carbon bond to a triple carbon-carbon bond.

2.1.2 Disulfide Bonds

One unique feature of proteins is the covalent bond between the sulfur atoms of two cysteine amino acid residues (the different amino acids will be discussed in detail in Section 2.2.1). Such disulfides stabilize a protein. They form in oxidizing environments. Normally, the cytosol is reducing. In eukaryotes disulfide bonds in proteins are formed in the oxidizing conditions of the rough endoplasmic environment or in the intermembrane space in mitochondria. Under reducing conditions these bonds can break and the sulfur atoms become protonated. The red-ox properties of such interactions are of great importance to certain proteins. The distance between the sulfur atoms in a disulfide bond is around 2 Å.

2.1.3 Charge Interactions

Electrostatic interactions are very important in biological systems. It is not possible to cover electrostatic theory here, and we will just mention a few points. If we look at a globular protein or at the surface of a lipid membrane we see that charged groups, such as lysine or arginine residues in proteins and phosphatidylethanolamine or phosphatidylserine groups on lipids (Chapter 6), are normally solvated and have nearby counterions. Not only would burying an isolated ionic group inside the protein or in the core of the lipid bilayer cause a loss of the solvation energy, but there would also be an electrostatic price to pay. The attraction between two oppositely charged ions is given by Coulomb's law, being proportional to $e^2/\epsilon r$, where e is the unit charge, r is the distance between the ions considered as point charges, and ε is the dielectric constant of the medium in which they are located. In aqueous solution ε is roughly 80, whereas it is much lower inside the hydrophobic core of proteins or the lipid bilayer where it is about 2-20. Thus, if a charge is buried, there is an enormous energetic advantage in burying a suitable opposite charge as close as possible. This is one of the reasons for the formation of saltbridges, for example, and why we need channels in membranes to transport ions through the membrane (Chapter 13).

Concerning the interactions and binding of ions to charged surfaces on proteins or membranes, the theory is more sophisticated. This pertains also to situations where we consider interactions between, for example, two membranes of the same or different charge or even one neutral and one charged membrane. Osmotic pressures and entropic forces will be involved in such events, and it is outside the scope of this book.

2.1.4 Hydrogen Bonds

Hydrogen bonds are central in biology. They contribute by stabilizing and orienting chemical groups with regard to each other. A proton interacting with two adjacent electronegative atoms with lone electron pairs (called the donor and acceptor) can form a hydrogen bond. These terms indicate that the proton can for some fraction of time be bound to either atom. The direction and the distance between the hydrogen bond donor and acceptor vary. The strength of a hydrogen bond is primarily related to its length and linearity. Normally hydrogen bonds are around 2.8 Å in length and the ΔH (change in enthalpy) of forming the bond is around 20 kJ/mol. The angle between the donor, the hydrogen and the acceptor is normally close to 180° . Deviations will decrease the strength of the bond. The angle at the acceptor atom (for example, the angle C-O-H when a hydrogen bond to a carbonyl oxygen is formed) is often 120° , corresponding to the position of a lone pair of electrons in the oxygen atom. Deviations in this angle appear to be less important. For example, in the hydrogen bonds stabilizing secondary structure in proteins, this angle is often close to linear, probably due to steric constraints.

In macromolecules the hydrogen bond donors and acceptors are normally nitrogen and oxygen atoms, where the donor has a covalently bound hydrogen and the acceptor has a free electron pair. Hydrogen bonds to sulfur atoms such as in cysteine are also found.

A macromolecule contains many hydrogen bond donors as well as hydrogen bond acceptors. In the interior of proteins or nucleic acid molecules these acceptors essentially always find a partner. The high energy of an unsatisfied bond in the interior is unfavorable. The tumor suppressor protein, p53, is an example of a highly unstable protein. Some hydrogen bonds in its interior lacking proper partners could contribute to its low stability.

2.1.5 van der Waals' Interactions

While the main chain of a protein is highly polar, many side chains are non-polar or hydrophobic. To the extent possible, they interact with each other rather than with the polar solvent, water. This has the effect that the hydrophobic side chains are normally found in the interior of proteins in van der Waals´ interactions with each other at distances between neighboring atoms of around 3.6 Å. The many such interactions in a biological macromolecule contribute significantly to its stability (see Section 3.1.1.1).

2.2 Amino Acids and the Protein Backbone

2.2.1 The Amino Acids

Proteins are composed of amino acids, which are made up of a tetrahedral α -carbon with four substituents: an amino group, a carboxyl group, a variable side chain and a hydrogen

atom (Figure 2.1). Because all four α -carbon substituents are different, amino acids possess an inherent chirality. This means that their mirror images are unique molecules that are not superimposable on each other (Figure 2.1). The one exception to this rule is the amino acid glycine, whose side chain consists only of one hydrogen atom (Figure 2.2).

All naturally occurring proteins are made up of amino acids in the L-configuration (where L stands for levo = left). The opposite hand (mirror image) is called the D-configuration (D stands for *dextro* = right). Viewing the L-conformation of an α -carbon down the H-Cα bond, one sees in clockwise orientation the carboxyl (COO⁻), the side chain (R) and the amino group (NH_3^+) (the CORN rule; Figure 2.1).

The sequence of the amino acids is the *primary structure* of a protein. In evolution, 20 amino acids were selected as the standard set (Figure 2.2). Two additional amino acids (selenocysteine and pyrrolysine) can also be genetically encoded but are rarely occurring. Numerous post-translational modifications of the amino acids have been identified (Table 2.2). Many of these have important signaling functions.

The amino acids can be divided into groups according to their properties. One way to classify them is the following:

Non-polar residues: Ala, Val, Ile, Leu, Met, Pro and Phe Charged polar residues: Asp, Glu, His, Lys and Arg Uncharged polar residues: Ser, Thr, Cys, Asn, Gln, Tyr and Trp No side chain: Gly

This classification is not always useful: the large aromatic residues Tyr and Trp as well as Cys are primarily hydrophobic although they can form hydrogen bonds. His is positively charged below a pH of around 7 and Pro has special properties (Figure 2.7 and Section 4.6.4.2) since its side chain connects to the main chain nitrogen.

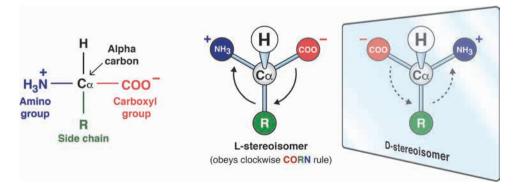


Fig. 2.1 ■ *Left*: The general structure of an amino acid. *Middle*: Illustration of the CORN rule for the L-stereoisomer of an amino acid. The α -carbon is viewed down the H-C α bond and the COO $^-$, R and NH₃ groups occur in the clockwise orientation. *Right*: An amino acid with the D-stereoisomer, the mirror image of the L-stereoisomer.

The amino and carboxyl groups of amino acids have pK_as of around 9 and 2, respectively. This means that both groups will usually be charged at neutral pH and the amino acid will exist in its zwitterionic form (i.e. it will have no net overall charge; Figure 2.1). The charge of the main chain termini as well as the amino acid side chains can be of functional relevance.

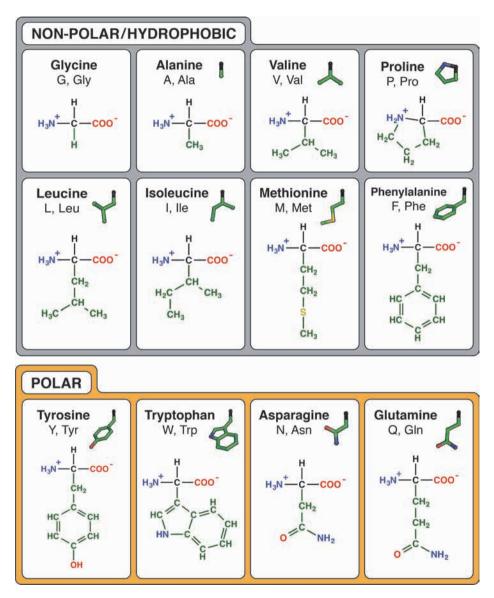


Fig. 2.2 ■ The 20 different amino acid side chains. In the stick drawing of the side chains the α -carbon is in black.

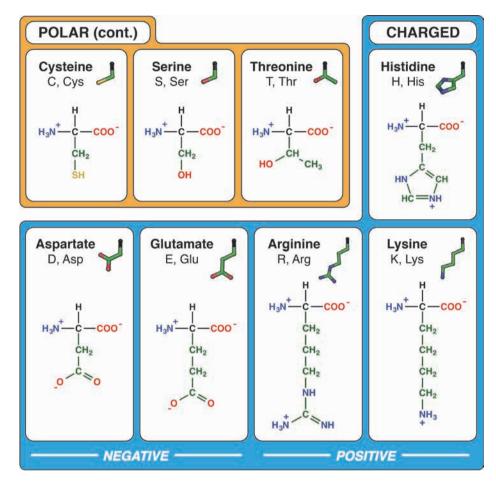


Fig. 2.2 ■ (Continued)

2.2.2 Side Chains and Their Interactions

The sequence of amino acid side chains gives proteins unique properties. They not only determine the fold of the protein, but they also determine surface properties that are important for selective interactions with other molecules and catalysis of chemical reactions.

The pK_a s of the amino acid side chains are given in Table 2.2. Arginine, lysine, aspartate and glutamate are normally charged. In addition, the side chains of cysteine, histidine, serine, threonine and tyrosine can sometimes be charged in physiological environments.

Depending on the environment, the pK_a s of amino acids (and hence their protonation states) can be subject to dramatic changes. For example, if two carboxyl groups (from Asp

TABLE 2.2 The Amino Acids

| Name | 3-Letter Code | 1-Letter Code | pK_a of Side Chain | Examples of Enzymatic Post-Translational Modifications (see textbox in Chapter 10) |
|---------------|------------------|------------------|----------------------|---|
| Alanine | Ala | A | _ | |
| Arginine | Arg | R | 12.5 | Methyl |
| Asparagine | Asn | N | | Glycosyl |
| Aspartate | Asp | D | 3.9 | |
| Cysteine | Cys | C | 8.2 | Disulfide Cys-Cys, prenyl |
| Glutamate | Glu | E | 4.1 | Carboxyl |
| Glutamine | Gln | Q | _ | |
| Glycine | Gly | G | _ | |
| Histidine | His | Н | 6.0, 14.5 | PO_4 |
| Isoleucine | Ile | I | _ | |
| Leucine | Leu | L | _ | |
| Lysine | Lys | K | 10.5 | Methyl, acetyl, carboxyl, ubiquitinyl, SUMOyl |
| Methionine | Met | M | _ | |
| Phenylalanine | Phe | F | _ | |
| Proline | Pro | P | _ | Hydroxylation as in collagen |
| Serine | Ser | S | 14.2 | Glycosyl, PO ₄ |
| Threonine | Thr | T | 15 | Glycosyl, PO ₄ |
| Tryptophan | Trp | W | _ | |
| Tyrosine | Tyr | Y | 10.5 | SO_4 , PO_4 |
| Valine | Val | V | _ | |

or Glu residues) are close to each other without any positively charged group to balance their negative charges, their pK_as may be raised considerably. Thus they will be protonated more readily to remove the repulsive force between two negatively charged groups.

The side chains can interact with each other or with the main chain in many different ways. The non-polar or hydrophobic side chains, mostly found in the inner part of the protein, interact with other non-polar side chains (Section 2.1.5). The polar side chains can hydrogen bond to each other or to the main chain (Section 2.1.4). The charged groups frequently interact with side chains of opposite charge on the surface of the protein forming salt bridges (Section 2.1.3).

Histidine can have several protonation states. Both nitrogens of the imidazole side chain can be protonated or deprotonated. An imidazole with one proton on either of the nitrogens is neutral in charge. When both nitrogens are protonated the imidazole is positively charged. In rare cases when an imidazole bridges between two metal ions with its

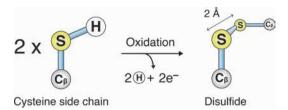


Fig. 2.3 \blacksquare The formation of a disulfide bond also showing its preferred conformation with a 90° angle between the S-C_B bonds viewed down the S-S bond.

two nitrogen atoms it is fully deprotonated. This is the case in copper-zinc superoxide dismutase where the imidazole moiety of a histidine residue bridges between the metals. In this case the net charge of the histidine side chain is −1.

Disulfide bonds have sometimes been artificially introduced into proteins by mutations to increase stability or to probe the flexibility of proteins. Here it has been proven necessary to obey the principle that is observed for naturally occurring S-S bonds. The C_{β} -atoms of the two cysteines involved have to be displaced by 90° when looking down the S-S bond (Figure 2.3).

There are many kinds of modifications of amino acid residues, especially lysine residues (Table 2.2). The enzymatic or non-enzymatic modifications of some amino acid side chains in vivo are further described in the textbox in Chapter 10.

2.2.3 Aromatic Interactions

The aromatic groups in nucleic acids (the bases; see Chapter 5) or in proteins (the side chains of Trp, Tyr, Phe and His residues; see Section 2.2.1) can provide a special type of stabilizing interaction. The cloud of π -electrons on both sides of the rings constitute a fractional negative charge while around the edge of the aromatic groups there is a corresponding fractional positive charge (Figure 2.4). These partial charges can stabilize interactions between aromatic groups in two main ways. The aromatic groups can stack onto each other. Here the stack is not a perfect pile of aromatic groups, but they are shifted in such a manner that the positive charge at the edge of one aromatic group interacts with the π -electrons of a neighboring aromatic group (Figure 2.4b). An alternative way is a perpendicular orientation between the interacting aromatic groups. Here again the edge of one ring faces the surface of another (Figure 2.4c). Finally, aromatic groups in proteins can interact with cationic species, such as the side chains of arginine and lysine residues, in cation- π interactions (Figure 2.4d). In addition, the cloud of π -electrons from several aromatic groups can interact with the hydroxyl-free sides of sugar residues (see Section 7.1), metal ions or methylated amino groups (for an example, see Section 10.3.1.2) and partly neutralize them.

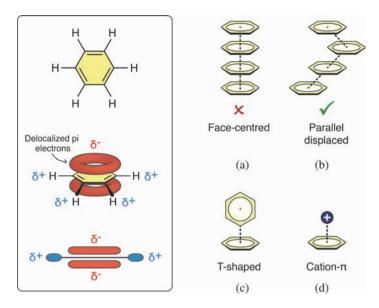


Fig. 2.4 • Aromatic interactions. *Left:* The arrangement of electrons in benzene. *Right:* (a) Stacking of aromatic groups on top of each other is unfavorable. (b) The stacking is always with a sideways displacement for optimal charge interactions. (c) Aromatic groups can also interact favorably in an orthogonal manner, so-called T-stacking. (d) The π -electrons of an aromatic group can generally interact with any positively charged moiety.

2.2.4 The Protein Backbone

In proteins, the amino acid residues are linked by peptide bonds (also known as *amide bonds*). These are covalent bonds between the carbonyl carbon from one amino acid and the amino nitrogen from the next amino acid (Figure 2.5). Peptide bond formation is catalyzed by the large ribosomal subunit (Chapter 11) and involves the liberation of one water molecule. In proteins and peptides, the free amino group of the first amino acid is called the *N-terminus*, and the free carboxyl group of the last amino acid is called the *C-terminus*. It is convention that protein sequences are always written from *N-terminus* to *C-terminus*, which is the order in which the protein is synthesized on the ribosome.

The peptide bond between the CO and NH groups has a partial double bond character due to a resonance between the major form (\sim 60%) and a form with a double bond between the C and N (\sim 40%; Figure 2.6). Rotation about a double bond is not possible, and therefore the six atoms between two consecutive C α atoms always lie in a plane called the *peptide plane*.

Given the constraints of the peptide plane, there are only two possible orientations of the protein backbone around the peptide bond: (1) the *trans* configuration, in which consecutive Cα atoms lie on opposite sides of the peptide bond to each other (Figure 2.7, top), or (2) the *cis* configuration, in which they lie on the same side (Figure 2.7, bottom). These two conformations are described by the *torsion angle* (or *dihedral angle*) of the peptide

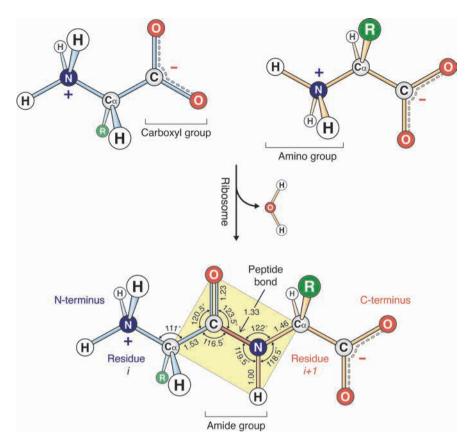


Fig. 2.5 ■ Peptide bond formation from two amino acids. The amino acids are shown here in their free zwitterionic forms, but when bound to tRNA on the ribosome both the amino and carboxyl groups are uncharged. The distances (in Å) and angles between the atoms of a peptide bond are shown in the bottom panel.

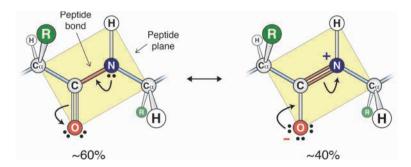


Fig. 2.6 ■ The peptide bond is a resonance between two electronic states that create the peptide plane. The protein main chain is colored blue and red (the peptide bond).

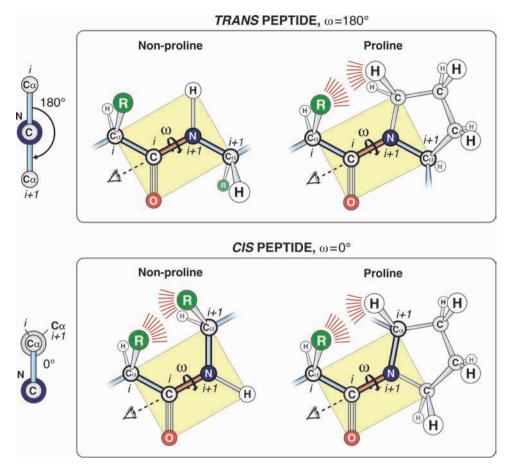


Fig. 2.7 ■ The *cis* and *trans* conformations of peptide bonds preceding non-proline and proline residues. Clashes are represented by red lines. For amino acids in general, the *trans* configuration is preferred due to the too close approach between two neighboring side chains in *cis*. For proline some conflict exists in both conformations.

bond, ω . When looking along the peptide bond, ω is the angle between the C α -C bond from residue i and the N-C α bond from residue i+1 (Figure 2.7). Therefore, the *trans* configuration will have $\omega = 180^{\circ}$, and the cis configuration $\omega = 0^{\circ}$.

cis peptide bonds are highly energetically unfavorable due to steric clashes between adjacent side chains (Figure 2.7), and therefore almost all peptide bonds in proteins exist in the *trans* configuration. When *cis* bonds do occur, special enzymes are often required to convert the bond from a *trans* to a *cis* configuration (Section 12.1.1.1).

Interestingly, *cis* peptide bonds are much more common if they precede proline residues (they represent around 6% of X-Pro bonds, compared with 0.04% of non-proline peptide bonds). This is because the cyclic proline side chain will create some steric hindrance even in the *trans* form (Figure 2.7). Therefore, the energy difference between the *cis* and *trans* configurations is diminished and *cis* bonds occur with higher frequency.

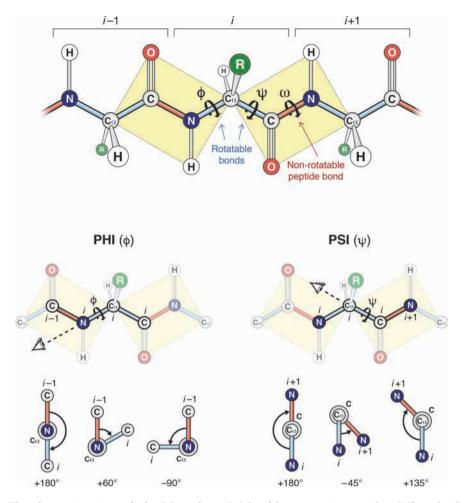


Fig. 2.8 ■ The determination of phi (φ) and psi (ψ) backbone torsion angles. When looking down the N_i - $C\alpha_i$ bond, ϕ is the angle between the C_{i-1} - N_i bond and the $C\alpha_i$ - C_i bond. When looking down the $C\alpha_i$ - C_i bond, ψ is the angle between the N_i - $C\alpha_i$ bond and the C_i - N_{i+1} bond.

Because of the planarity of the peptide bond, the conformation of the polypeptide backbone can be fully described by just two torsion angles per residue (Figure 2.8). These are the angles around the only two rotatable bonds in the protein backbone: the N_i-Cα_i bond (measured by the angle phi, ϕ) and the $C\alpha_i$ - C_i bond (measured by the angle psi, ψ). The method for calculating ϕ and ψ angles is shown in Figure 2.8. ϕ and ψ lie between –180° and +180°, and they have a positive value if — when looking down the N_i -C α_i bond (for ϕ) or the $C\alpha_i$ - C_i bond (for ψ) — the rear bond is rotated clockwise relative to the front bond.

The ϕ and ψ angles have restricted values due to steric clashes between the carbonyl oxygen, the hydrogen of the NH group, the hydrogen on the Cα carbon and the side chain atoms. This was used to define the allowed regions as described by the Ramachandran plot (Figure 2.9). The two main allowed regions in the Ramachandran plot correspond to the two main types of conformation (α-helices

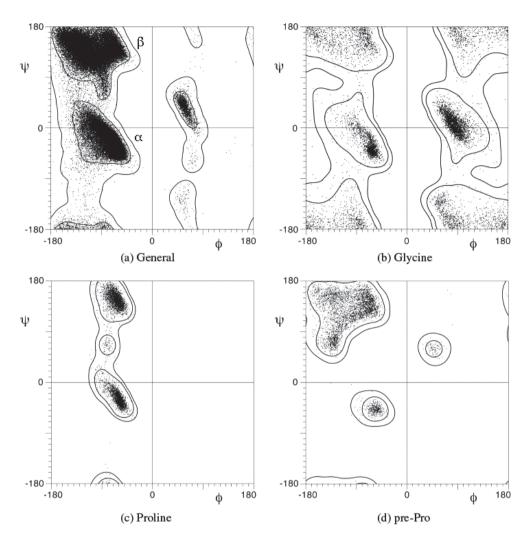


Fig. 2.9 Ramachandran plot: The observed Ramachandran angles for general amino acid residues (a) and glycine (b). The highly populated areas on the left-hand side of each plot correspond to the β-structure (top) and right-handed α-helix (below). The highly populated area on the right-hand side corresponds to the left-handed α-helical conformation. The observed Ramachandran angles for proline residues (c) and residues preceding prolines (d). (Reprinted with permission from Lovell *et al.* (2003), Structure validation by $C\alpha$ geometry: ϕ , ψ and $C\beta$ deviation. *Proteins* **50**: 437-450. Copyright (2003) Wiley.)

and β -sheets) observed in proteins. A small region corresponding to a left-handed helical conformation is also allowed (see Section 2.3.1.2 for a discussion of right- and left-handed helices). The allowed regions for Gly residues are much larger since there is no side chain to restrict the angles. On the other hand, proline residues have a very restricted set of conformations with ϕ restricted to values close to -60° . The Ramachandran plots

in Figure 2.9 are based on conformational angles observed in protein structures and differ somewhat from the theoretical description.

When the distribution of the conformational angles for high quality protein structures is plotted, the β region has two separate maxima. The plot for glycine residues shows a quite different pattern, with a strong peak at the region for left-handed helices.

Ramachandran plots can be used to analyze the overall quality of a protein model obtained from experiments. For a well-determined structure the great majority of the conformational angles are in the preferred regions. Some individual residues in proteins might still be found in less favored or even disallowed regions. In some cases these unfavorable conformations are connected with biologically important properties of the protein. (A discussion about the quality of protein structures is found in the text about the Protein Data Bank, Section 19.1.2.2)

2.2.5 Side Chain Conformations

Two of the amino acids have chiral β -carbons, namely, isoleucine and threonine. Only one of the stereoisomers is found in naturally occurring proteins, and this stereoisomer is defined by the "CARC rule" (Figure 2.10). The CARC rule is analogous to the CORN rule described in Section 2.2.1.

For steric reasons, most side chains also have preferred conformations. Each tetrahedral carbon atom of the side chains has four substituents, and will prefer a staggered conformation where the substituents are as far as possible from the substituents on the adjacent atoms (Figure 2.11). This leads to three alternative torsion angles (chi, χ) with 120° intervals. Figure 2.11 shows how χ angles are calculated, using the amino acid glutamate as an example. Analysis of protein structures shows that amino acid residues have strong preferences for certain combinations of angles. These conformations, rotamers, can be collected into libraries that are important in protein modeling (Table 2.3).

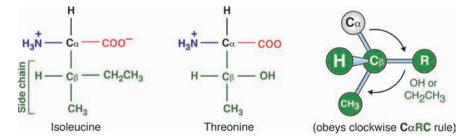


Fig. 2.10 ■ The configuration of the isoleucine and threonine side chains in normal proteins viewed down the H-Cβ bond (the CARC rule).

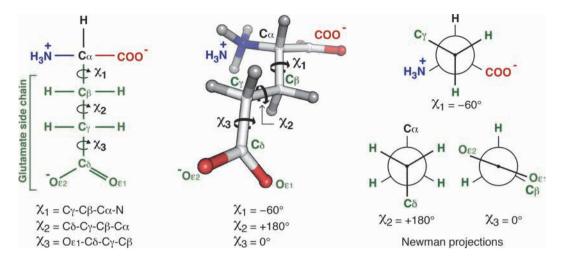


Fig. 2.11 ■ Illustration of the torsion angles in a side chain. *Left*: The definition of the angles in the side chain of a glutamate residue. *Middle*: A common rotamer for a Glu side chain. *Right*: Newman projections looking down the Cβ-Cα bond (χ_1 , top), the Cγ-Cβ bond (χ_2 , bottom left), and the Cδ-Cγ bond (χ_3 , bottom right). χ angles have a positive value if the rear bond is rotated clockwise relative to the front bond. Negative values indicate anticlockwise rotation. The other staggered conformations correspond to a rotation of 120° in the positive or negative direction.

TABLE 2.3 Preferred Rotamers for Four Amino Acid Residues

| Residue | Rotamer | χ1 | χ2 | χ3 | Frequency |
|---------|---------|------|------|-----|-----------|
| GLU | -t0 | -70 | -177 | -11 | 0.27 |
| GLU | tt0 | -176 | 175 | -7 | 0.26 |
| GLU | 0 | -65 | -69 | -33 | 0.11 |
| GLU | -+0 | -56 | 77 | 25 | 0.09 |
| GLU | +t0 | 70 | -179 | 7 | 0.09 |
| GLU | t+0 | -174 | 71 | 14 | 0.06 |
| GLU | +-0 | 63 | -80 | 16 | 0.05 |
| HIS | | -63 | -74 | | 0.34 |
| HIS | t- | -175 | -88 | | 0.25 |
| HIS | -+ | -70 | 96 | | 0.16 |
| HIS | +- | 68 | -81 | | 0.14 |
| HIS | t+ | -177 | 101 | | 0.09 |
| HIS | ++ | 48 | 86 | | 0.02 |
| ILE | -t | -61 | 169 | | 0.45 |
| ILE | | -60 | -64 | | 0.18 |

(Continued)

| Residue | Rotamer | χ1 | χ2 | χ3 | Frequency |
|---------|---------|------|-----|----|-----------|
| ILE | +t | 62 | 164 | | 0.16 |
| ILE | tt | -167 | 166 | | 0.13 |
| ILE | t+ | -175 | 72 | | 0.03 |
| VAL | t | 174 | | | 0.67 |
| VAL | _ | -63 | | | 0.26 |
| VAL | + | 69 | | | 0.05 |

Val has a single torsion angle, His and Ile have two and Glu has three. The actual value for the angles is the average of observed values for this type of rotamer. The preferred staggered conformations are trans (180°), gauche + (60°) and gauche — (-60°), denoted t, + and — in the table. The observed terminal χ angles for glutamates (i.e. χ_3), as well as glutamines and some other residues, deviate from the staggered conformations. In the table this is denoted with 0. From Ponder JW, Richards FM. (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes, *J Mol Biol* 193: 775–791.

2.3 Secondary Structure

The Ramachandran plot shows that proteins have a limited variability of main chain conformations. Frequently a stretch of amino acids adopts the same conformation. This leads to what is called the *secondary structure* of proteins. The protein backbone has two main types of regular secondary structure, corresponding to the two main regions in the Ramachandran plot (Figure 2.9). These are the *alpha* (α)-*helix* and the *beta* (β)-*sheet*, formed by several β *strands*. Helices and β -sheets are efficient ways to satisfy the need of hydrogen bond donors and acceptors in the interior of proteins. Parts of the protein without any regular pattern of conformational angles are called *loop* or *coil regions*. These are normally found on the surface of proteins. These regions are sometimes flexible.

2.3.1 Helices

2.3.1.1 The α -helix

A polypeptide can adopt a number of different helical conformations. In practice only three are regularly occurring (see Table 2.4). The most common type is called the α or 3.6_{13} helix. In this nomenclature the number 3.6 indicates the number of residues per turn and

the subscript (13) indicates the number of atoms in the ring that is formed by the hydrogen bond between main chain NH and CO groups. In α -helices, the protein backbone forms hydrogen bonds between carbonyl oxygens and NH hydrogens in the next turn (residues n and n+4). There is a rise of 1.5 Å per residue and the pitch, the rise per turn of the helix, is 5.4 Å (Figure 2.12). The ideal backbone ϕ and ψ angles of α -helices are about –60° and –45°, respectively (Figure 2.9 and Table 2.4).

TABLE 2.4 Parameters Describing Helices in Proteins

| Type | 3 ₁₀ | α | π |
|---|-----------------|-------|-----------|
| Residues per turn | 3.0 | 3.6 | 4.1 |
| Atoms in H-bonded ring | 10 | 13 | 16 |
| Hydrogen bonding | n-n+3 | n-n+4 | n - n + 5 |
| Angle between neighboring residues | 120 | 100 | 88 |
| Helical rise per amino acid residue (Å) | 2.0 | 1.5 | 1.15 |
| φ (°) | -71 | -60 | -75 |
| ψ (°) | -18 | -45 | -40 |

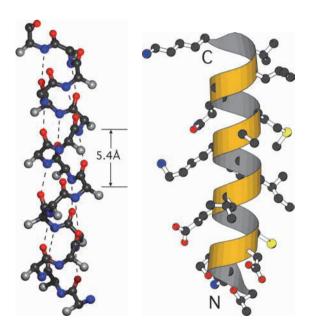


Fig. 2.12 ■ The α -helix. *Left*: The main chain and C β atoms (grey) of an α -helix. The pitch (rise per turn) is 5.4 Å. *Right*: The same α -helix showing the side chains. The backbone is drawn schematically, with the C β atoms pointing towards the N-terminus of the helix (down).

The side chains will point out from the helix axis at intervals of about 100° (since there are 3.6 residues in one complete 360° turn), and the Cβ atoms are directed towards the N-terminus of the helix.

Different amino acids have different propensities to be located in an α -helix. In particular, proline and glycine have low propensity for α -helices. If proline occurs in a helix it will bend or kink the helix since the cyclic side chain will disrupt the helical hydrogen bonds (see Figure 4.14). On the other hand, prolines are frequently found at the aminoterminal end of helices where they are not involved in the hydrogen bonding of the helix.

2.3.1.2 The Handedness of Helices

All helices or twists possess an inherent chirality or "handedness", and can be either righthanded or left-handed. This concept arises frequently in structural biology, and is not only used to describe the 3D configuration of α -helices, but also DNA (Section 5.2), β -sheets (Section 2.3.2) and coiled-coils (Section 3.3.2). One way to determine the handedness of a helix is to look down its axis and trace along the backbone in the clockwise direction. If the path traced is going into the page, then the coil is right-handed (Figure 2.13). If it is moving out of the page, then it is left-handed.

Almost all α -helices are right-handed. This is because all naturally occurring proteins are built from L-amino acids (Section 2.2.1), which cannot readily form a left-handed helix due to steric clashes with the side chains (the opposite would be true if proteins were instead built from D-amino acids). However, although rare, some short left-handed helices (~4 residues long) have been identified in protein structures. These usually contain at least one glycine residue, which has greater backbone flexibility due to the small side chain (Figures 2.2 and 2.9).

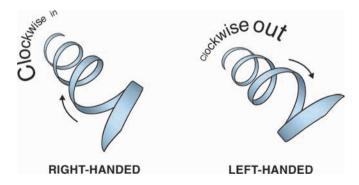


Fig. 2.13 \blacksquare Determining the handedness of an α -helix. Right- and left-handed helices are nonsuperimposable mirror images of each other. Note that the helix can be viewed from either end and the handedness will not change — it is an inherent property of the helix.

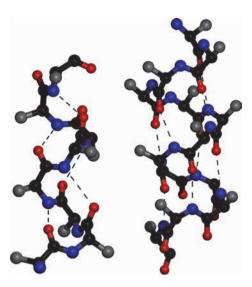


Fig. 2.14 ■ A 3_{10} helix from *Aplysia limacina* myoglobin (PDB: 1MBA) and a π -helix from methane monooxygenase hydroxylase from *Methylococcus capsulatus* (PDB: 1MTY).

2.3.1.3 The 3_{10} and π -helices

There are other types of helices in addition to the α -type helix. Most common is the 3_{10} helix, where the hydrogen bonds are formed between residues n and n + 3. The 3_{10} helix is narrower and not as stable as the α -helix, and there are few examples of long 3_{10} helices (Figure 2.14).

The other type of helix is called the π -helix. It has hydrogen bonds between residues n and n+5. This is the least frequent form of helix. This type of helix has been classified as 4.4_{16} . However, an analysis of π helices shows that the proper number of residues per turn is approximately 4.1 and the name 4.1_{16} would thus be more appropriate (Table 2.4). The π -helix has been considered quite unstable due to the larger circumference that would leave a hole down the center of the helix. However, when the π -helices are analyzed there is no significant hole along the helix due to the tilting of the peptide planes along the helix.

Normally a few residues at the ends of helices can adopt the hydrogen bonding pattern of π -helices, but in quite a number of proteins at least seven consecutive residues with the π -helix conformation have been observed.

2.3.1.4 Helical Dipoles

Helices have one important element in common. Since all the peptide planes are oriented in essentially the same way and they can all be considered as small dipoles (that is, all δ - CO groups point in one direction and all δ + NH groups point in the other direction; Figure 2.12), the helices become larger dipoles. In effect, the amino end of a helix has a

partial positive charge and the carboxyl end has a partial negative charge. Negatively charged side chains at the N-terminus and positively charged residues at the C-terminus frequently stabilize helices. Likewise the binding of charged ligands such as coenzymes or substrates can be strengthened by the partial charges at the ends of helices. Due to electrostatic interactions, neighboring and antiparallel helices also stabilize each other.

2.3.1.5 Polyproline Helices and Collagen

Proline-rich sequences can adopt a special type of secondary structure called the polyproline helix (Figure 2.15). There are two variants called polyproline I (PPI) which is much denser and based on prolines with cis conformation. PPI helices are right-handed. Polyproline II (PPII) is more common and is based on the trans conformation of proline, and these helices are left-handed. In the Ramachandran plot, the PPI conformation is found in the β structure region around $\phi = -75^{\circ}$ and $\psi = 160^{\circ}$ and PPII at $\phi = -75^{\circ}$ and $\psi = 150^{\circ}$. PPII is a frequently occurring conformation in coil regions but not always is identified as polyproline II. Spectroscopically PPII has also been observed in natively unfolded proteins.

PPII is the dominating type of secondary structure in collagen. Here, three polypeptide chains with the highly repetitive sequence (proline — hydroxyproline — glycine)_n each with a left-handed PPII helical conformation wrap around each other and form a special type of triple-stranded superhelix which is right-handed (Figure 2.15). This type of structure is also found in rare cases in soluble proteins such as C1q in the complement system and the glycine-rich N-terminal domain of ObgE, a factor involved in translation on the ribosome.

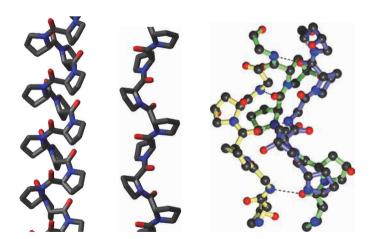


Fig. 2.15 ■ Left: The structures of the two helices polyprolines I and II (from Wikipedia). Right: The triple helix of a segment of a collagen. The three chains have a repeated sequence of prolinehydroxyproline-glycine. The glycine residues allow the chains to come in close contact and form inter-chain hydrogen bonds (PDB: 1Q7D).

2.3.2 Beta Structures

2.3.2.1 β -sheets

The other type of common secondary structure is the β -sheet. These are formed by stretches of extended polypeptide chains, called β strands, where the CO and NH groups can form hydrogen bonds to neighboring strands on both sides. β -sheets can be parallel, antiparallel or mixed, depending on the direction of the strands (Figure 2.16). In an ideal sheet, all NH and CO groups of internal strands form hydrogen bonds with CO and NH groups of neighboring strands. Some sheets form closed barrels (Section 4.7.2), but in open sheets, the strands at the edges will have free NH and CO groups.

Along the polypeptide chain, the side chains extend from the sheet on alternating sides. Neighboring strands in the β -sheet have neighboring side chains pointing in the same direction — the side chains form lines perpendicular to the chain direction (Figures 2.16 and 2.17). The β -sheets are not flat but always twisted in the same direction, though the degree of twist differs between different sheets (Figure 2.17).

 β -strands at the end of a sheet, with a number of CO- and NH-groups exposed to solvent have a tendency to bind further β strands and extend the sheet. This is seen in numerous proteins where subunit aggregation is accomplished through pairing of

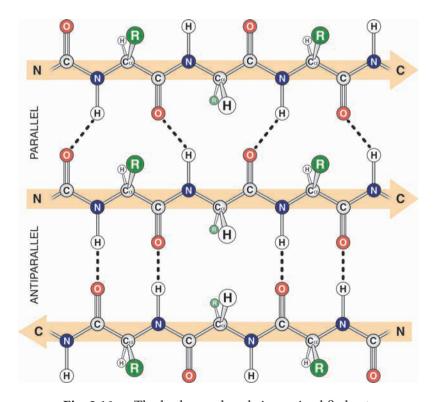


Fig. 2.16 ■ The hydrogen bonds in a mixed β -sheet.

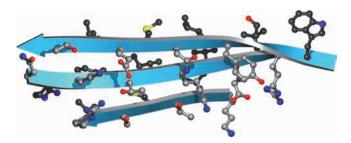


Fig. 2.17 \blacksquare An antiparallel β -sheet showing the positions of the side chains on both sides of the sheet. Along the strands the side chains point alternatingly in opposite directions, but across the strands neighboring side chains point in the same direction. The side chains on the lower side of the sheet have been drawn with grey carbon atoms and those on the upper side are shown in black (PDB: 2BU1). β-sheets are always twisted to some degree.

β-sheets. In the same way, peptide substrates bind to proteolytic proteins through a β-interaction between substrate and enzyme. The amyloid aggregation of proteins also occurs through extensive β strand interactions of some part of a protein that would normally be integrated into a folded structure and may be in conformations different from the aggregated form (see Section 3.3.3).

$2.3.2.2 \beta$ -bulges

In antiparallel β-structures bulges are regularly observed. This occurs when there is an additional amino acid residue in one strand that shifts the residues out of phase with one another compared to the ideal hydrogen-bonding pattern (Figure 2.18). These β-bulges give the β-sheets an increased twist. β-bulges have a few standard conformations and could be included among the regular secondary structures.

2.3.2.3 Turns

The reverse turn is another type of regular secondary structure. They are short and often connect two β strands. The reverse or β -turn ideally has a hydrogen bond between the CO of residue n and NH of residue n+3 (Figure 2.19). This type of tight turn imposes strong restrictions on the conformational angles of residues n+1 and n+2 (Figure 2.19). There are a few types of such turns and in some of them there is a strong preference for certain residues (Gly or Pro) in specific positions where the torsion angle properties of these residues are important for the stability of the turn (Table 2.5).

Another type of turn found in many proteins is the γ-turn, where a hydrogen bond is formed between the CO group of residue n and the NH hydrogen of residue n+2.

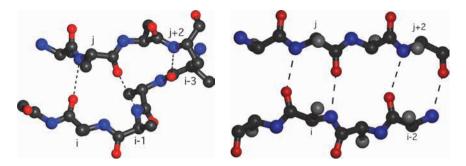


Fig. 2.18 ■ *Left*: A β -bulge in an antiparallel β -sheet: the NH and the CO groups of residue j in the upper chain form hydrogen bonds with the CO group of one residue i and the NH group of the next i–1 residue in the lower chain. Hydrogen bonds are also formed between residues j+2 and i–3. *Right*: For comparison, the normal hydrogen bond pattern in an antiparallel sheet is shown. The NH and CO groups of residues i and j, and between i–2 and j+2, are formed.

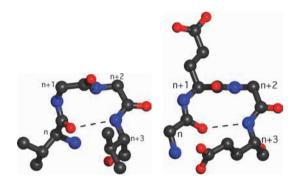


Fig. 2.19 ■ Two of the most common types of reverse turns, type I' and II.

| TABLE 2.5 | Conformational | Angles and | Preferred | Amino | Acid |
|--------------|----------------|---------------|------------|---------|------|
| Residues for | the Most Comm | on Types of l | Reverse or | β-Turns | |

| Type | Phi | Psi | Phi | Psi | Amino Acid Preferences |
|------|------|------|------|-----|------------------------|
| I | -60 | -30 | -90 | 0 | Pro (n+1), Gly (n+3) |
| I′ | 60 | 30 | 90 | 0 | Gly (n+2) |
| II | -60 | 120 | 80 | 0 | Pro (n+1), Gly (n+2) |
| II' | 60 | -120 | -80 | 0 | Gly (n+1) |
| VIa | -60 | 120 | -90 | 0 | |
| VIb | -135 | 135 | -75 | 0 | |
| VIII | -60 | -30 | -120 | 120 | |

[&]quot;n" is the first residue of the turn. The conformational angles are for residues n+1 and n+2.

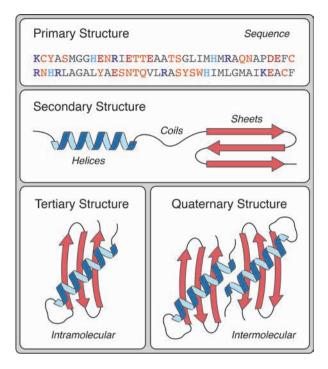


Fig. 2.20 ■ A pictorial illustration of the four different levels of protein structure.

Tertiary and Quaternary Structure

The tertiary structure of a protein (Figure 2.20) is the conformation or fold that the protein has when it is biologically active. This is also called the native conformation. Many proteins form stable oligomers with other copies of the same protein or with other proteins. This is called the *quaternary structure* of the protein. Both these levels of structural organization are discussed in the following chapter.

For Further Reading

Original Articles

Fleming PJ, Rose GD. (2005) Do all backbone polar groups in proteins form hydrogen bonds? Prot Sci 14: 1911-1917.

Fodje MN, Al-Karadaghi S. (2002) Occurrence, conformational features and amino acid propensities for the π -helix. *Prot Engin* **15**: 353–358.

Lovell SC, Word JM, Richardson JS, Richardson DC. (2000) The penultimate rotamer library. *Prot Struct Funct Genet* **40:** 389–408.

Richardson JS, Getzoff ED, Richardson DC. (1978) The β bulge: A common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci USA* **75**: 2574–2578.

Reviews

Adzhubei AA, Sternberg MJE, Makarov AA. (2013) Polyproline-II helix in proteins: Structure and function. *J Mol Biol* **425**: 2100–2132.

Chothia, C. (1984) Principles that determine the structure of proteins. *Ann Rev Biochem* **53**: 537–572.

Steiner T. (2002) The hydrogen bond in the solid state. *Angew Chem Int Ed* 41: 48–76.

The Folding, Folds and Functions of Proteins

3.1 Protein Stability and Dynamics

3.1.1 Hydrophobic Effect

The stability of the folded protein is due to a number of factors, one of which is the formation of a *hydrophobic core* that is of utmost importance. This core is formed by the packing of non-polar side chains in the interior of a protein or a protein domain. The hydrophobic core is primarily formed because of the hydrophobic effect that is entropic in nature. In the unfolded state, hydrophobic side chains in a protein are exposed to the water solvent, but in the folded state, a majority of these side chains are shielded from the solvent, leading to an increase in the entropy of the water molecules. If polar groups are present in the hydrophobic core, they further stabilize the protein structure by hydrogen bonds or salt bridges.

The polar groups in the main chain of the polypeptide normally form regular patterns of hydrogen bonds in the core such as β -sheets and α -helices. Since it is energetically unfavorable to bury polar groups in the hydrophobic core, helices and β -sheets form the central parts of most proteins where all the polar groups of the protein backbone are involved in hydrogen bonds. More information about the hydrophobic effect is given in Chapter 6 about lipid membranes.

3.1.1.1 The folding process

The conformation or tertiary structure of a protein is defined by its amino acid sequence and therefore by the nucleotide sequence of the corresponding gene. The direct relation between the conformation and the amino acid sequence of a protein was first proven by Anfinsen's experiments, where a functional staphylococcal nuclease was refolded from denatured fragments of the protein. We now know a lot more about the process by which the chain of amino acid residues finds its correct conformation, the *folding process*.

There are an enormous number of theoretically possible conformations of the polypeptide chain of even a small protein. Levinthal made a classical statement that for a protein of 100 amino acid residues it would require more time than the age of the universe to sample three conformations of each Ramachandran angle, a sum of 3¹⁹⁸ conformations. Nevertheless, the folding of proteins is fast, which means that the polypeptide chain cannot sample all these conformations to find the most stable one. A possible explanation for this phenomenon is that the folding follows one or a limited number of pathways to the final fold. A different way to express this is that folding occurs in an energy landscape, which can be described as a funnel. The protein may find the absolute minimum or get trapped in some local minimum (Figure 3.1). Folding occurs while the polypeptide chain is growing, so-called co-translational folding. It cannot be excluded that the folding process itself and its kinetics are important for the final conformation, and that not only the folded (native) conformation, but also the folding process are coded in the amino acid sequence.

For most proteins, the native conformation seems to be the most stable conformation, although the energy difference between the folded and unfolded state in general is small (around 20–40 kJ/mol). The forces that make a protein fold are no different from those that lead to the formation of secondary structure: covalent forces, hydrogen bonds, electrostatic and van der Waals' interactions. In addition, the hydrophobic effect makes a critical contribution. Hydrogen bonds (see Section 2.1.4) are a central aspect of protein folding. They provide an essential contribution to the energetics of protein folding, but in addition they are directional and cooperative and determine to a large extent the fold of

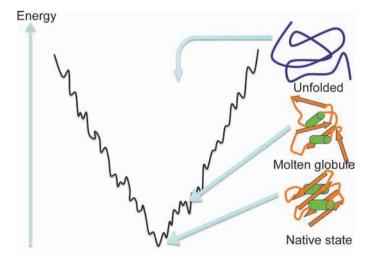


Fig. 3.1 ■ The folding funnel — an illustration of the folding process of a protein.

a protein. The loss of a couple of hydrogen bonds would be enough to prevent the protein from folding.

It is possible that the native fold does not represent the conformation of lowest energy since the folding process might not allow a pathway reaching this state. In such a case the native state is the conformation of lowest energy that can be reached by the folding process. The small difference in energy between the unfolded and folded state is also a property that is important for the biological function of proteins and can lead to different types of flexibility. Furthermore, the control of various biological processes requires that proteins have a limited lifetime and can be degraded. A very stable protein conformation is not consistent with these requirements.

The folding pathway seems to differ considerably between proteins. Some proteins seem to fold without any intermediate stages. For other proteins, the folding involves intermediates, sometimes called molten globule states, where some of the native interactions are established, but the fully compacted conformation has not been reached. For example, secondary structure elements may already have formed, but are not yet optimally packed in these intermediates.

In cells, there are a number of proteins, chaperones that aid some proteins in their folding (see Chapter 12). Chaperones are not specific for a certain fold. They bind nonnatively folded proteins in a way that prevents them from aggregating but allows them to refold. Their role is thus to improve the efficiency of the folding rather than to guide the folding process to a specific conformation.

Some proteins can only fold properly as precursor proteins, where a segment of the polypeptide chain is removed after the folding is complete. A well-known example is the small protein-hormone insulin. The function of such a segment is probably to remove a kinetic barrier in the folding pathway.

3.1.1.2 Packing of proteins and empty holes in proteins

For the stability of a protein, the packing of secondary elements is of great importance. Water molecules associated with unfolded polypeptides are mostly released during the formation of the protein fold. This gives a gain in entropy, stabilizing the protein. Some water molecules remain trapped in the structure where the hydrogen bonds can be satisfied (see Section 3.1.2.2). However, some smaller cavities or packing defects remain empty in the structure. Many of those are too small to bind anything. In other cases, the empty holes are larger and hydrophobic, between 25 Å³ and 150 Å³. Noble gases such as xenon or krypton can be made to bind to these larger cavities.

The loss of stability of a protein due to such cavities has been studied in the case of lysozyme from the bacteriophage T4. A range of mutants of the hydrophobic core were examined. Smaller side chains, such as alanine, replaced larger hydrophobic ones. These mutations mostly did not change the overall structure of the protein, although the residues around the generated cavity to some extent moved into the cavity. No water molecules were observed to bind in the hydrophobic cavities. The loss of stability could be estimated at 100 to 140 J/mol/Å^3 of cavity volume or 80 J/mol/Å^2 of cavity surface area.

The structure of a protein is a primary determinant of its function but dynamics also plays a crucial role. For a protein to be able to adopt different conformations there is a need for parts of the protein to move into new positions. These conformational differences may be small but still important for the function. If there are empty spaces in the protein this will allow certain parts of it to adopt different conformations. Thus, empty holes in a protein are of functional significance. This has been extensively investigated for myoglobin by different techniques and using a range of site-specific mutations. Transport of ligands to and from the heme group uses cavities in the protein, and the binding of CO in these pockets can be followed by spectroscopic techniques as well as time-resolved crystallographic studies (Laue crystallography) after photodissociation. Studies of mutants that influence these cavities show that the rate of transport depends on the packing in the core. This example illustrates that the amino acid sequence of a protein not only codes for the fold of a protein but implicitly also affects its dynamic properties.

3.1.1.3 Denaturation — thermal stability

Proteins are marginally stable. The stability of the folded state can be described as the difference between two large numbers with opposite signs, the enthalpy (H) and the entropy (S). The free energy describing the stability of a protein is given by

$$\Delta G = \Delta H - T\Delta S$$

where T is the temperature. Proteins can easily be denatured in various ways, by adding denaturing chemicals (urea, guanidinium hydrochloride), change of pH or by increasing the temperature. Some proteins are stabilized by cross-linking by S-S bridges or binding of metal ions.

Comparisons of proteins from different organisms with the same function but different amino acid sequence show that they can have very different degrees of stability against denaturation, which may reflect the different conditions in which these organisms live. Thus, the enzymes of thermophilic ("heat-loving") organisms are generally more stable to thermal denaturation. On the other hand, they are less active than their corresponding mesophilic enzymes at low temperatures. The latter in turn are less active than their psychrophilic ("cold-loving") counterparts.

Understanding thermal denaturation is an important prerequisite for the use of enzymes at elevated temperatures in industrial processes. How can the thermal stability be improved for enzymes? From mutational analysis it has been found that cavities in proteins are destabilizing. Thus, filling such cavities with bigger side chains has been beneficial. Another stabilizing factor is to increase the number of hydrogen bonds

and charge-charge interactions. The introduction of S-S bridges can also increase the stability.

3.1.2 Water Molecules

3.1.2.1 Bound water

Most protein molecules are folded and function in an aqueous environment. Their relationship with water is a central issue in discussing protein structure and function. Protein surfaces, not involved in binding to other proteins, are mainly hydrophilic, but with hydrophobic regions. In structural studies of macromolecules numerous water molecules have been identified on the surface, in pockets and internally in the structures. The higher the resolution (see Section 3.2) of the crystal structure, the better the water molecules can be identified. The water molecules in the first hydration shell are integral parts of the protein and contribute to its stability and function. The water molecules on the surface of a protein that can be identified crystallographically are normally hydrogen bonded to one or several polar groups on the protein. In some places, parts of a second layer of water molecules can be seen.

The oxygen atom of the water molecule has tetrahedral geometry with two hydrogens and the two free electron pairs (lone pairs). The water molecule can thus act as a donor of two hydrogen bonds and acceptor for another two. The distances between the donor and the acceptor atoms in these hydrogen bonds vary between 2.7 Å and 3.3 Å. In specific environments (e.g. when bound to a metal ion) the pKa of the water molecule may be perturbed and the pH, at which the system is studied, may be such that specific water molecules may be deprotonated or protonated to OH⁻ or H₃O⁺, respectively.

3.1.2.2 Internal water molecules

Water molecules that are completely buried inside a protein molecule are also frequently observed. Certain proteins have water filled channels partly or completely through the protein. In the case of membrane transporter proteins (Chapter 13), these channels may be part of functional pathways for transport of water, protons or ions across a membrane. Sometimes the water molecules are extensively hydrogen bonded, but water molecules forming small clusters can be stabilized at least transiently even in weakly polar or nonpolar cavities. The exchange of internal water molecules can be studied by NMR spectroscopy or neutron diffraction.

Water molecules can be regarded as trapped during protein folding. A partly polar cavity between groups of the protein would prefer a water molecule rather than an empty space in which hydrogen-bonding potential is not fulfilled. Some water molecules of this type nevertheless exchange very rapidly with external water molecules as has been observed by NMR. Protein molecules go through "breathing" motions that simplify the exchange of these water molecules.

3.2 Tertiary Structure: Protein Folds

Proteins can occur in numerous shapes. The classical image of a protein is a globular structure. The first protein structures that were determined were those of myoglobin, hemoglobin and hen egg white lysozyme and they were quite globular. However, the structures of elongated or fibrous proteins have been known for a long time. Keratin in hair is a distinctly fibrous protein made of α -helices. Silk, on the other hand, is a fibrous protein made up of β -strands. The known protein structures are stored in the Protein Data Bank, PDB (See Chapter 19).

3.2.1 Domains

Many proteins are formed around a single hydrophobic core, but the majority of proteins are built of separate folding units, *domains*. A domain is normally formed around a separate hydrophobic core, and this can be used as a definition of a domain. Usually, the domains are separate units along the polypeptide chain. One class of domain arrangement in proteins is formed through gene duplications and fusion. The same fold can in this way be repeated once or several times along the polypeptide chain. A classical case is that of the aspartyl proteases, which are sometimes built up of two identical subunits related by a two-fold axis. However, in other cases these enzymes are composed of one polypeptide with two domains that are structurally homologous and related by an approximate two-fold axis.

In many cases there is no well-defined separation between parts of proteins. The classification of a protein part as a domain or a sub-domain then becomes subjective. Although domains are separate folding units, they are not always continuous along the polypeptide chain. Several segments of the polypeptide extend from one domain to form a separate domain. In this case, there are at least two connections between the domains.

3.2.2 Classification of Protein Folds

The secondary structure elements and the way they are connected form the protein *topology* and the *tertiary structure* of the protein. There is an enormous variation in the

way proteins fold, and therefore it is useful to classify folds to allow comparisons between proteins. Two independent efforts to organize protein folds have resulted in the CATH and SCOP databases (Chapter 19). In each of them the basic unit for classification is a protein domain. When proteins are composed of several domains, and if these have different folds, they will end up in different places in the database. The two databases are both hierarchical, grouping protein domains according to the major secondary structure elements found (mainly α , mainly β or mixed α and β), and then subdividing these classes into several unique topologies. The enormous variability in natural amino acid sequences does not translate into a similar variability in folds. Many proteins have the same or similar folds, even in cases where they have no obvious evolutionary relationship. An important aspect of these databases is that the protein domains are classified into superfamilies with a presumed evolutionary relationship. These classifications are mainly based on sequence similarity, complemented by structural and functional similarities.

3.2.3 Topologies and Motifs

3.2.3.1 Antiparallel and parallel β -sheets

An analysis of existing proteins shows that there are some arrangements of secondary structure elements that are more common than others. Such arrangements are called motifs. Several architectures formed by β -strands are shown in Figure 3.1. Antiparallel β-sheets are often formed by β-hairpins (two consecutive antiparallel β-strands linked by a loop or turn). If the hairpin motif is repeated we get an up-and-down sheet called B-meander.

Up-and-down sheets are found in many proteins. In many cases, they form an open sheet, but they can also form barrels or cylinders (Figure 3.2). A large group of proteins are the β-propeller proteins, which often have blades of four-stranded antiparallel up-and-down sheets. There are examples of four-, five-, six-, seven- and eight-bladed propellers and a single example of a ten-bladed propeller. In most of these proteins, one of the propeller blades is formed by the N-terminal and the three C-terminal strands, in this way closing and locking the propeller structure.

A four-stranded antiparallel sheet can also be formed with the Greek key topology. This motif can be seen as a hairpin that is bent to produce four strands. Greek key motifs are also very common in proteins containing β -sheets.

A special type of a Greek key structure is an extended version called jellyroll. This is an eight-stranded arrangement that forms a β sandwich of two four-stranded.

Proteins mainly composed of β structure normally have antiparallel sheets, but one exception is proteins with β -helix or β -solenoid topology, where the chain forms short β-strands in a parallel and helical manner into a triangular prism.

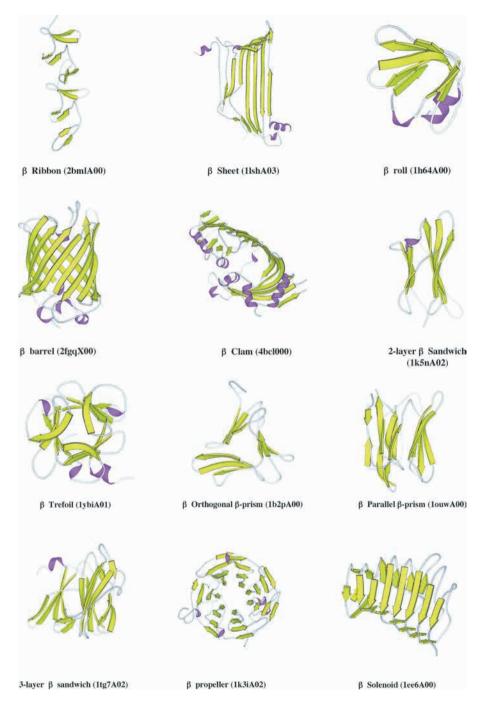


Fig. 3.2 ■ Architectures in the β class in CATH (v 3.1).

3.2.3.2 Folds with combinations of helices and strands

A large number of folds combine helices and strands (Figure 3.3). The observed architectures often show a sheet with parallel strands. Several of these, like the common $\alpha\beta$ barrel (often called TIM barrel) and 3-layer $\alpha\beta\alpha$ sandwich is formed by a motif, the $\beta\alpha\beta$ unit, formed by two parallel strands and a connecting helix. This motif exists in two forms, but

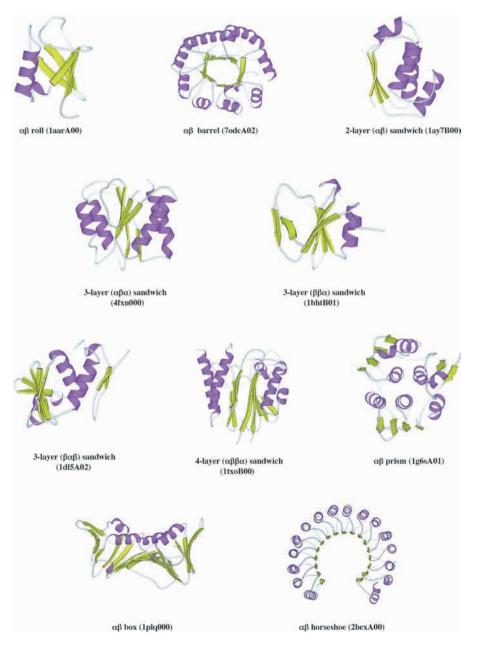


Fig. 3.3 ■ Architectures with α and β structure.

one of them, the right-handed version, is by far the most common. The TIM barrel has got its name from the enzyme, triose phosphate isomerase. This fold is formed by eight β - α units, $(\alpha/\beta)_8$ generating a cylinder. The eight parallel strands form a closed cylinder, and the eight helices make a layer outside the β cylinder. This fold is found in a large number of enzymes. Sometimes one pair of β -strand and α -helix is deleted making the structure an open $(\alpha/\beta)_7$ structure. The Rossmann fold, found in many proteins, is an example of a topology with the 3-layer $\alpha\beta\alpha$ sandwich architecture. It is built of such $\beta\alpha\beta$ motifs. It was first found in lactate, maleate and alcohol dehydrogenase, where a six-stranded parallel sheet is at the core of the protein. In the Rossmann fold, the N-terminal strand is in the middle of the sheet. Two $\beta\alpha\beta$ units form half of the sheet ($\beta\alpha\beta\alpha\beta$), and the rest is formed by a similar unit starting next to the first strand and related by an approximate two-fold axis. Because of the handedness of the units and the two-fold relationship, the helices will end up on opposite sides of the sheet. There are many proteins with similar topology, but with slight variation in the number and order of the strands.

3.2.3.3 Helix packing

There are only a few architectures found in helical proteins (Figure 3.4). The packing of helices tends to be either parallel or orthogonal. A common type of helix packing is the four-helix bundle, where two pairs of antiparallel helices are arranged with their helix axes at about 20°. Four-helix bundles are found in many proteins. A simple form is the

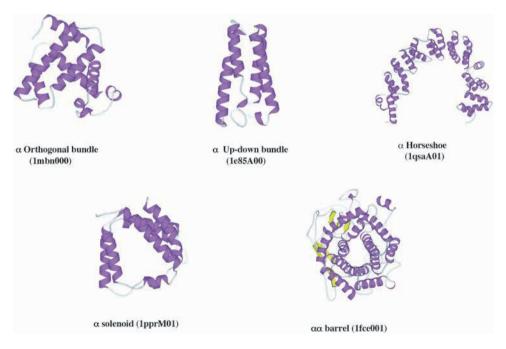


Fig. 3.4 • Architectures in the α class.

up-and-down bundle. The orthogonal bundle, as in the globin fold, is an example of the other type of helix packing.

3.2.4 Flexibility and Disorder in Proteins

3.2.4.1 Loops and tails

Secondary structure elements are connected by turns or loops. Such loops can sometimes be quite long and protrude from the more compact part of the protein. The loops can be devoid of secondary structure or they can be organized into a β -ribbon or a helix. Other extensions from a compact structure are amino-terminal or carboxy-terminal tails. These tails can also have a secondary structure. Both loops and tails usually have structural roles, quite frequently for the stabilization of oligomeric structures. One example is lactate dehydrogenase where the N-terminal 20 amino acid residues form an arm that folds onto another subunit to stabilize the interaction between dimers in the tetrameric enzyme (Figure 3.5). The related enzyme malate dehydrogenase lacks this N-terminal extension and is therefore often dimeric.

Coat proteins of viruses (Chapter 18), histones (Chapter 9) and ribosomal proteins (Chapter 11) often also have extended positively charged parts that neutralize the negative charges of the nucleic acids. These extended elements may be the most ancient protein elements of these large aggregates and possibly of proteins in general. Compared to metal ions they can have the allosteric advantage of having several positive charges and

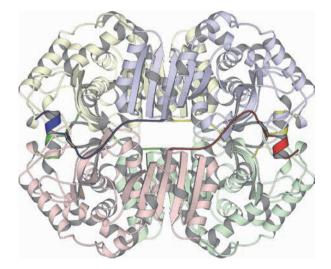


Fig. 3.5 ■ The tetrameric enzyme lactate dehydrogenase highlighting the N-terminal extended arm that interacts with other subunits keeping the complex together (PDB: 1LDM). The related enzyme malate dehydrogenase from some species lacks this extended arm and is therefore often dimeric. The dimer corresponds to the blue-green or red-yellow interaction.

can bridge between several phosphates in a nucleic acid. The globular domains may be later additions to these peptides giving higher specificity and added functions.

3.2.4.2 Natively unfolded proteins

Most proteins are probably ordered although they are not rigid, but in many cases small or large portions of proteins have a high degree of flexibility. The tails discussed above are in many cases disordered in the isolated proteins. One explanation for this is that the function of the flexible portion is to bind to other molecules, but this portion is not stable in the absence of the binding partner. Furthermore, the bound conformation might be extended or have a shape that makes it topologically impossible to bind as a rigid body.

There is an increasing awareness that there is a large fraction of proteins that are natively unfolded or are transiently folded. An example of an intrinsically unstructured protein (IUP) is thymosin $\beta 4$, which binds to G-actin and prevents assembly of fibrous actin (Section 15.1). In the bound form, helices bind to opposite ends of the actin monomer preventing its aggregation. The protein uses helices as binding elements. These helices exist temporarily in the isolated unfolded protein, thereby reducing the entropic cost of binding a completely unstructured protein to its target.

3.2.5 Convergence and Divergence

3.2.5.1 Common folds

The existing folds are the result of their evolutionary history, rather than the suitability of a certain type of fold for a certain type of function. Still, many functions are connected to specific types of folds. One example is immunoglobulin: all immunoglobulins have several domains, all of them β sandwiches with very similar topology. These types of domains are not only found in the proteins of our immune system but also in many cell surface receptors. The cause of this similarity is probably that most or all of these proteins have diverged from a common origin.

In addition to the immunoglobulin fold, there are a number of other folds that are found in many seemingly unrelated proteins. A number of common folds are listed in Table 3.1.

Many enzymes have the TIM barrel fold. Many functionally unrelated enzymes with no obvious sequence similarities have this fold, and it appears that this fold has developed independently in the evolution of these proteins. The active sites of these enzymes are always found at the same end of the barrel.

Another common type of fold is the "jellyroll", which is found in numerous viral coat proteins as well as in a number of other proteins with completely unrelated functions, such as tumor necrosis factor. Other folds have been found only in a single family of proteins

TABLE 3.1 A Number of Common folds Found in Different Protein Families

| Name | Type of Fold | Examples of Fold | Figure |
|-------------------------------|---|--|------------|
| α/β doubly wound | Mostly parallel sheet with helices on both sides | Ras, subtilisin, adenylate kinase | 3.6, 11.18 |
| TIM barrel | Cylinder of eight β-strands interconnected by helices | Triose phosphate isomerase, glycolate oxidase, aldolase | 3.3 |
| Split α/β sandwich | Antiparallel sheet with helices on one side | 4Fe-4S ferredoxin, acyl phosphatase, RNA binding proteins | 3.3, 3.18 |
| Immunoglobulin | β sandwich | Immunoglobulin, receptor domains, superoxide dismutase | 17.1 |
| α up-and-down | Four pairwise antiparallel helices | Hemerythrin, TMV coat protein | 3.4 |
| Globin | Two layers of non-parallel helices | Hemoglobin, phycocyanin | 3.4 |
| Jellyroll | β sandwich | Tumor necrosis factor, viral coat proteins, concanavalin A | 18.5 |
| Trefoil | Cylinder formed by three sheets | Interleukins, ricin | 3.2 |
| Ubiquitin αβ roll | Small sheet with helices on one side | Ubiquitin, 2Fe-2S ferredoxin | 12.29 |

with a unique function. Two examples are the enzyme carbonic anhydrase and the coat protein of small RNA bacteriophages.

Many proteins have similar folds with a simple topology. Whether these similar folds in functionally unrelated proteins have evolved by divergence is unknown. Such simple folds might have been favored during evolution because of the possibilities of an efficient and accurate folding process.

As the structure is determined for more and more proteins, the number of observed folds is increasing steadily. The enormous amounts of data coming out of the efforts to sequence complete genomes give us a possibility to estimate the number of existing folds. This is done by trying to classify all sequences into groups of homologous proteins. Since the conformation is more conserved than the sequence, every protein in such a group will have the same fold (as long as the assignment is correct!). Many proteins with the same fold may be evolutionarily related but have diverged so much that no obvious sequence similarity remains, while other proteins may have arrived at the same fold independently and be truly non-homologous. From the proportion of new sequences that belong to already known groups one can extrapolate to make an estimate of the total number of groups of homologous proteins with significant sequence similarity that exist in living cells. From the number of folds that have been connected to distinct sequence groups, we can estimate how many of these groups will have a unique fold. The present estimate is that there are at most a few thousand different folds. Of these, about 1000 are known.

The large number of currently known folds is not the total number of folds that can be formed by a polypeptide (sometimes called *fold space*). The total number of possible amino acid sequences is enormous. For a small protein of 100 amino acids, there are 20¹⁰⁰ possible sequences. Probably, only an extremely small fraction of all theoretically possible sequences would be able to fold into a stable globular structure, since the relatively rigid backbone and the requirement to form a reasonably well packed hydrophobic core severely restrict the folding. It is possible that there are many stable folds that do not exist in nature, since the existing folds are the result of an evolutionary process that has not explored all these sequence combinations.

3.2.5.2 Serine proteases: a case of functional convergence

It is important to note that the classification of protein folds in structural classes is not linked to any functional classification. In general, a certain type of function is not restricted to a certain class of folds. This is illustrated by the serine proteases and related enzymes, where there are several distinct families of proteins with similar function but with different folds. The serine protease chymotrypsin belongs to a large group of mostly eukaryotic proteolytic enzymes. It is composed of two similar domains that have a β barrel fold (Figure 3.6). The active site and catalytic mechanism of chymotrypsin is very

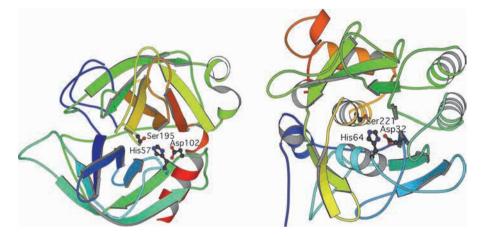


Fig. 3.6 ■ The folds of the serine proteases chymotrypsin (*left*) and subtilisin (*right*) are very different, but the proteins have the same function and enzymatic mechanism. The Ser-His-Asp catalytic triad is indicated (PDB: 4CHA and 1GCI, respectively).

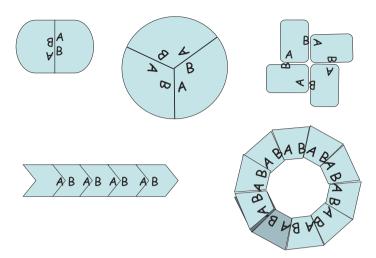


Fig. 3.7 ■ Some examples of protein oligomers. Depending on the relative orientation of the surfaces that interact with each other (A and B) the oligomer could become a dimer, a trimer, a tetramer, a linear (open) arrangement or a helical oligomer.

similar to that of subtilisin, which is a bacterial enzyme with a completely different fold (class $\alpha + \beta$). Obviously, chymotrypsin and subtilisin do not have a common origin. However, the three residues involved in the catalytic mechanism (the catalytic triad) Ser-His-Asp are the same but occur in a different order in the sequence. Nevertheless the catalytic center and the mechanism are quite similar. This is a case of *convergent evolution*: the same function has evolved independently in two proteins. The serine protease example shows that the actual fold of a protein can be regarded as a means to form a stable structural scaffold. On this scaffold, active sites and other functional properties are molded. Another case is the enzyme carbonic anhydrase with several different folds for the same function (see Chapter 8).

3.3 Quaternary Structure: Formation of Protein **Oligomers**

The function of biological macromolecules depends not only on their structures, but also on their interactions. Not only smaller molecules and ions but also the interactions between the macromolecules to a large extent determine their function. Proteins can form large oligomeric aggregates composed of only one type of building block (homo-oligomers) or several types (hetero-oligomers). To be able to form a stable interaction the loss of accessible surface area needs to be of a significant size. An average size for stable complexes is around 1600 $Å^2$ (which is the sum of the interface areas from both chains). These interactions are

of the same nature as in the folding of proteins, but generally hydrophobic residues and the hydrophobic effect play a central role. Naturally charges or hydrogen bonds need to find balancing partners. Proteins have also evolved to avoid unsuitable interactions or aggregation that may damage its interactions. A positive aspect of protein aggregation is the *in vitro* crystallization of proteins for structural determination.

3.3.1 Oligomeric Proteins, Symmetry and Breaking of Symmetry

The variety of arrangements in oligomeric proteins seems unlimited. When identical protein molecules interact the same interaction surfaces of the different protein subunits (*protomers*) are normally used. This can lead to a closed arrangement between a fixed number of molecules or to an open interaction between many molecules. The closed arrangements usually have some type of symmetry (Figure 3.7, Table 3.2).

Open interactions could be linear (less likely) or helical. A number of well-characterized structures have helical symmetry. Some examples of helical aggregation are listed in Table 3.3.

The closed objects could have one or several symmetry axes. A simple dimer having one two-fold axis is the simplest form of an oligomeric protein. There are examples in biological systems of up to 39-fold rotational symmetry (Figure 3.8, Table 3.4).

The most complex symmetry of oligomers includes multiple symmetry axes with different directions. The simplest arrangement has tetrameric symmetry with three perpendicular

TABLE 3.2 Cyclic Symmetry Arrangements in Proteins

Rotational Symmetry Example

| Rotational Symmetry | Example | Comments |
|---------------------|--|---------------|
| 2 | Alcohol dehydrogenase | |
| 3 | Porin, Influenza virus hemagglutinin | |
| 4 | Influenza virus neuraminidase, aquaporin | |
| 5 | AB ₅ -enterotoxins | |
| 6 | C-phycocyanin | |
| 7 | GroES | |
| 8 | Light-harvesting complex 2 | |
| 9 | Light-harvesting complex 2 | |
| 11 | trp RNA-binding attenuation protein (TRAP) | |
| 16 | Light-harvesting complex 1 | |
| 17 | TMV coat protein ring | A double ring |
| 39 | Vault ribonucleoprotein | |

| TABLE 3.3 | Some Examples of Helical Aggregation of Proteins |
|-----------|--|
| | |

| Protein | Helical Rise Per Protomer (Å) | Rotation Angle Between Protomers (°) | Comments |
|----------------------------|----------------------------------|--|---|
| Actin | 27 | 167 | |
| Type IV pilus | 10.5 | 100 | |
| Filamentous phage, fd | 32 | 72 and 180 | Rings of pentamers with opposite direction. |
| Tobacco mosaic virus (TMV) | 1.4 | 22 | |

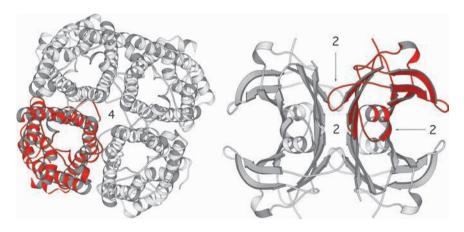


Fig. 3.8 ■ Two arrangements of tetramers. *Left*: Aquaporin (four-fold rotational symmetry, PDB: 3D9S). Each molecule has a channel for water molecules through its middle. Right: Transthyretin (222 symmetry, PDB: 1002).

TABLE 3.4 Protein Complexes with Multiple Symmetry Axes

| Type of Symmetry | Example | Comments |
|------------------|---------------------------------|---|
| 222 | Lactate dehydrogenase | 4 subunits |
| 32 | Aspartate carbamoyl transferase | 12 subunits of two kinds |
| 422 | Hemerythrin, glycolate oxidase | 8 subunits |
| 622 | Glutamine synthetase | 12 subunits |
| 72 | GroEL | 14 subunits |
| 23 | Protechuate 3,4-dioxygenase | Tetrahedral symmetry, 12 subunits |
| 39/2 | Vault ribonucleoprotein | 78 subunits |
| 432 | Ferritin | Octahedral symmetry, 24 subunits |
| 532 | Icosahedral viruses | Icosahedral symmetry, 60, or multiples of 60 subunits |

two-fold axes (222 symmetry, Figure 3.8). There are examples of three-, four-, six- and seven-fold symmetry combined with perpendicular two-fold axes. Most complex are the different forms of cubic symmetry, for example, the icosahedral viruses with 532 symmetry (Table 3.4). The cubic symmetry is generated by four three-fold symmetry axes along the space diagonals of a cube.

3.3.1.1 Polymerization - depolymerization

Several proteins can undergo polymerization - depolymerization reactions depending on the physiological need. An example is the muscle protein actin (see Section 15). This protein exists in two forms, G- and F-actin. F-actin is the aggregated or filamentous form. Depending on the interaction with a large number of proteins it can form helical aggregates or depolymerize. Actin forms thin filaments in muscle cells. Another protein, tubulin, can polymerize and depolymerize, controlled by other proteins to form the microtubules.

A number of virus coat proteins also undergo depolymerization upon infection of its host, and polymerization before they are ready to infect new cells.

3.3.1.2 Domain swap

Monomeric proteins composed of several domains can sometimes aggregate in a nonnative fashion through *domain swap* arrangements. Instead of interacting within the same polypeptide the domains may interact with the corresponding domain in another polypeptide chain (Figure 3.9).

Sometimes this can be done in a way that extends the oligomerization in a linear or possibly branched way. The final effect will be aggregation of monomers.

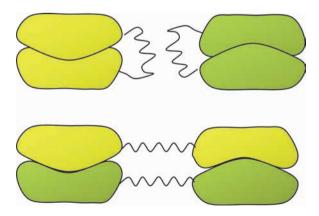


Fig. 3.9 ■ Domain swapping of a protein. Normally, the protein is a monomer (*top*) where the two domains interact within the monomer. *Bottom:* The two domains of the protein interact in the same way with a different monomer.

| Name | Repeat Motif | Structural Role | Functional Role |
|-------------|--|---|--|
| Leu-Zipper | Leu every 7 th residue, heptad repeat. | Hydrophobic interactions along the helix. | Dimerization of proteins |
| Coiled-coil | Heptad repeats (abcdefg) with a and d being hydrophobic. | Interactions between two or more helices. | Can form long "rope"- like structures (myosin, tropomyosin). |
| Gly-Zipper | GXXXGXXXG | Right-handed helix packing. Close approach between helices. | Membrane channel proteins |

TABLE 3.5 Helix Packing Motifs

3.3.2 Coiled-coils and Heptad Repeats

3.3.2.1 Overall features

In some proteins α-helices in subunits interact to form oligomers in ways that have identified them as special motifs. They are all basically coiled-coils. They have been given different and suitable names, e.g. Leu-zipper (Table 3.5).

One distinctive type of helix packing often exploited in protein-protein oligomerization is the coiled-coil, in which individual helices wind around each other to form extended superhelices. Classical coiled-coils are composed of between two to five helices arranged in a parallel or an antiparallel manner, but more complex ring-like arrangements with up to 12 helices are also found in nature.

Most coiled-coils are formed from a sequence motif called the heptad repeat, which consists of seven amino acids — designated abcdefg — that are repeated along the helices of the supercoil. The residues at positions a and d are almost always hydrophobic amino acids such as leucine, isoleucine or valine, which pack against each other to form a hydrophobic interface between the coiled helices. Burial of the hydrophobic a and d residues is the primary driving force for helix association, and the remaining *bcefg* positions are usually occupied by polar amino acids that interact with the aqueous solvent. Because of the importance of the a and d residues in coil formation, the heptad repeat is also sometimes referred to as a 3-4 repeat (since the a and d residues are spaced 3 and then 4 amino acids apart).

3.3.2.2 The molecular basis of coil formation

The molecular basis of coiled-coil formation lies in one of the fundamental properties of the α -helix that was described in Section 2.3.1.1: that each turn of an α -helix is made up of 3.6 residues. With 3.6 residues per turn, the seven residues of a heptad repeat are not quite enough to complete two full revolutions of the helix (7.2 residues). As a consequence, the a and d residues from adjacent heptad repeats will not line up along one face

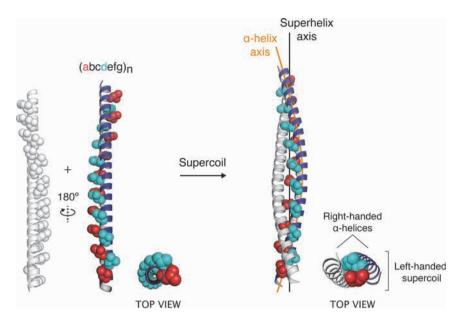


Fig. 3.10 ■ The heptad repeat and coiled-coil formation. The left panel shows two straight helices consisting of multiple consecutive heptad repeats. The side chains from a and d residues (all leucines for simplicity) are shown as spheres. For clarity, the a and d side chains are only shown for one of the two helices in the supercoil.

of a straight helix but will instead slowly rotate around its axis (Figure 3.10, left panel). It is thus impossible to bury all of the a and d residues by association of two straight helices.

To compensate for the systematic drift of *a* and *d* residues, the helix axes themselves must twist around each other in the opposite sense (or "handedness") of the right-handed helix backbone, forming a left-handed superhelix (Figure 3.10, right panel). The *a* and *d* amino acids spaced seven residues apart will then all point inwards towards the superhelix axis, creating a hydrophobic core between the two chains (Figure 3.11). Therefore, coiled-coil formation has effectively reduced the number of residues per turn of each helix from 3.6 to 3.5.

3.3.2.3 Side chain interactions at coiled-coil interfaces

The side chain interactions at coiled-coil interfaces are often represented using *helical wheel diagrams*, which show the position of each residue when the superhelix is viewed from above. Helical wheels are constructed by mapping the *abcdefg* heptad repeat — forming two turns of a coiled-coil helix — around the radius of two opposing circles (Figure 3.11).

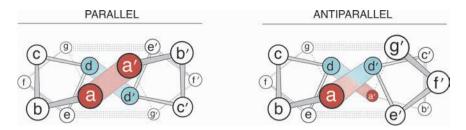


Fig. 3.11 ■ Interactions between interfacing residues in two-stranded parallel and antiparallel coiled-coils. In the helical wheel diagrams, dots indicate electrostatic or hydrogen bonding interactions involving polar e and g residues. Hydrophobic interactions between diagonally-opposite a and d residues are shaded red and blue.

From Figure 3.11, it is clear that the molecular details of the superhelix interface are different in parallel and antiparallel coiled-coils. When the helices are parallel, the a residues from one helix pack against diagonally-opposite a' residues from the other helix, and dresidues pack against d' residues. This forms repeating a-a' layers and d-d' layers along the length of the superhelix. In antiparallel coiled-coils the layers are mixed, with a residues interacting with d' residues and d with a'. Furthermore, parallel helices have interhelical e-g' contacts (usually salt bridges or hydrogen bonds) flanking the hydrophobic core, whereas antiparallel helices have g–g' polar contacts. The complementarity of these interactions confers specificity at the interface and often plays an important role in determining whether the superhelix is parallel or antiparallel.

The interfacing residues in all coiled-coils pack according to the "knobs-into-holes" principle that was first postulated by Francis Crick in the 1950s. In this packing mode, interfacing a and d residues protruding from one helix ("knobs") pack into cavities ("holes") on the opposing helix that are formed between four adjacent side chains. This concept is often visualized using helical net diagrams, which describe the relative positions of side chains in an α -helix in two dimensions. To understand these diagrams, Crick suggested that one should imagine circling a piece of paper around an α-helix and marking upon it the positions of the amino acid side chains (Figure 3.12). Unrolling the paper then produces the helical net, which can be used to model the geometry and interactions of a coiled-coil interface.

The knobs-into-holes packing patterns for two-stranded parallel and antiparallel coiled-coils are shown in Figure 3.12. For parallel coiled-coils, a' knobs pack into $d_{-1}g_{-1}ad$ holes and d' knobs pack into $adea_{+1}$ holes (where -1 and +1 refer to residues in preceding or following heptad repeats, respectively). This pattern is reversed in antiparallel coiledcoils. It is also evident from Figure 3.12 that, in order for the knobs from one helix to fit into the holes on the opposing helix, the helices in coiled-coils must be tilted at an angle of approximately 20° relative to each other. This same angle is also commonly observed between the α-helices of transmembrane proteins, which often interact via knobs-intoholes side chain packing (Section 4.6.3.1).

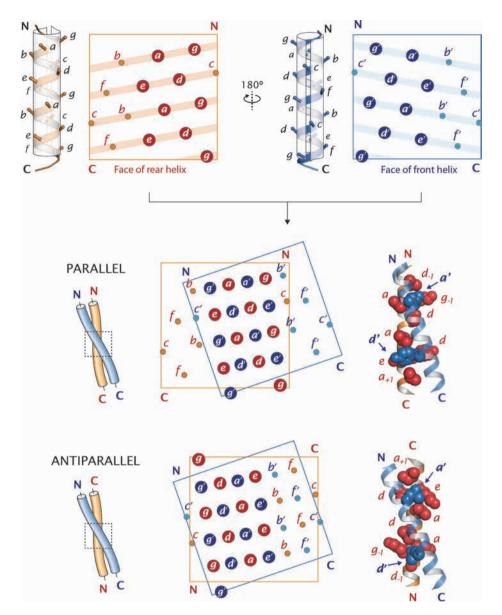


Fig. 3.12 ■ Helical net diagrams and knobs-into-holes packing. *Top:* The creation of helical net diagrams. The interfacing a, d, g and e residues are shown in large bold circles. *Bottom:* The knobs-into-holes packing principle illustrated by the superposition of helical nets in parallel and antiparallel orientations. The two helices cross at an angle of approximately 20° . The parallel coiled-coil structure is the protein tropomyosin (PDB: 1IC2), and the antiparallel structure is from the hepatitis D delta antigen (PDB: 1A92).

3.3.3 Amyloid Aggregation

In order to perform their functions, proteins have to arrive at the proper fold. The chaperone system is a system that prevents incorrect folding of proteins (see Chapter 12). Certain proteins have a tendency to fold in a manner that leads to aggregation into ordered insoluble fibers, called amyloids. These types of deposits can lead to numerous diseases (Table 3.6). A joint name for these problems is protein-misfolding disorders (PMDs).

Normally, the fibers consist of one single protein that has aggregated. There is no general relationship between the proteins that can form amyloids. However, structural investigations of the amyloid fibers show a unique and common fiber diffraction pattern, which is called cross-β (Figure 3.13). In the meridional direction (along the fibril) there is a peak corresponding to around 4.7–4.8 Å, which is characteristic of the strandto-strand distance in a β -sheet. This suggests that β -strands line up and form a β -sheet along the axis of the fibril. In the equatorial plane, the fiber diffraction pattern has a peak corresponding to a distance of 8-10 Å. This peak can be explained by the stacking of multiple β-sheets perpendicular to the fibril axis. Depending on the amino acid composition of the sheets, they will stack with somewhat different distances. Figure 3.14 shows a parallel and twisted β -sheet that would generate the cross- β diffraction pattern. Since all β-sheets are twisted (Section 2.5.4.4), the stacked sheets twist around each other along the fibril axis.

TABLE 3.6 Protein Misfolding Disorders

| Disease | Protein |
|-------------------------------------|----------------------------------|
| Alzheimer's disease | Aβ-peptide |
| Creutzfeldt-Jakob's disease | Prion protein |
| Parkinson's disease | α-Synuklein |
| Huntington's disease | Huntingtin |
| Type II diabetes | Islet amyloid polypeptide (IAPP) |
| Familial amyloidotic polyneuropathy | Transthyretin |
| Senile systemic amyloidosis | Transthyretin |
| Hereditary systemic amyloidosis | Lysozyme |
| Finnish-type familial amyloidosis | Gelsolin |
| Light-chain amyloidosis | Immunoglobulin VL domain |
| Familial British dementia | ABri |
| Aortic medial amyloid | Medin |
| Secondary systemic amyloidosis | Serum amyloid A |
| Spinocerebrellar ataxias | Ataxins 1,3,7 |
| Haemodialysis related A | β-2 microglobulin |

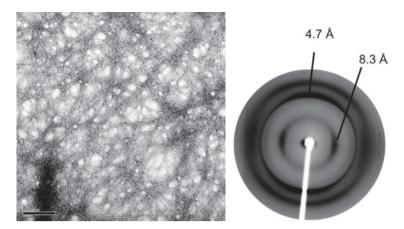


Fig. 3.13 • *Left*: Amyloid fibrils from IAPP (residues 1–37) (Kindly provided by G. Westermark.) *Right*: One example of the cross-β diffraction pattern of amyloid fibers. The 4.7 Å distance corresponds to the strand-to-stand distance within a β-sheet. The 8.3 Å reflection is variable and due to the distance between packed sheets (Reprinted with permission from Bader R, Seeliger MA, Kelly SE, *et al.* (2006) Folding and fibril formation of the cell cycle protein Cks1. *J Biol Chem* **281**: 18816–18824. (Copyright (2006) Elsevier.)



Fig. 3.14 ■ A peptide that has aggregated into a parallel β structure that would generate the cross- β diffraction pattern. Such β -sheets can then pack on each other in a number of ways.

If a protein containing a β structure has an edge that is not blocked by side chains or other structures it is free for extension of the β -sheet. This seems to be the case for transthyretin, one of the proteins that can lead to PMD. Sometimes the amyloid aggregation involves more than the β -strand interactions. If the β -strand that develops is part of a linker between two domains of the protein the domain can also become part of the aggregation in the form of runaway domain swapping (Figure 3.15).

Certain peptides have been identified to be mainly responsible for amyloid formation. The structures of some of them give detailed pictures of the organization of the cross-β pattern. The crystal structures of the peptides have the cross-β arrangement. The peptides form parallel or antiparallel β-strands that are perpendicular to the fiber axis (Figure 3.16). The distance between the strands is 4.9 Å. The amino acid residues of the peptides are in perfect register and not only the main chain but also the side chains make hydrogen bonds to one another. The β-sheets have two different interfaces with neighboring β-sheets. One is a dry interface where the distance between sheets is 8.5 Å. This dry interface is unusually tight with an extensive complementarity in the fit of the side chains between the sheets. The other, the wet interface, has a number of bound water molecules and the distance can be 15 Å. This may be a regular feature of the amyloid fibers.

Prion diseases such as transmissible spongiform encephalopathy (TSE) or "mad cow disease" are related to PMDs. The protein PrPC, particularly mutant forms, can aggregate. The aggregates are of the disease-causing form, PrPSc, which will induce soluble PrPC to change conformation to PrPSc and aggregate. These prion proteins are similar to the amyloid-forming proteins in the way they aggregate by forming extensive β structures. Prions are infectious because aggregated material from one individual can induce aggregation in a healthy individual. However, it is an open question whether amyloids in general are infectious.

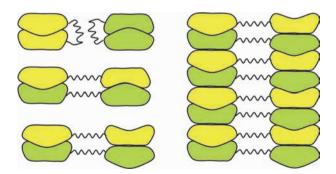


Fig. 3.15 ■ An illustration of the runaway domain swap phenomenon. In the upper left, two monomers are shown. They can form swapped dimers in two different ways (lower left). At the same time, the connection between the domains forms a β -sheet. These two ways of forming dimers enables aggregation of the protein into fibrils through both β-sheet and swapped domain interactions.

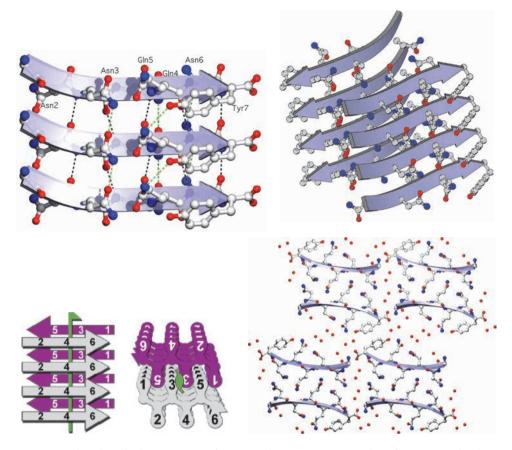


Fig. 3.16 ■ *Top*: The detailed structure of a peptide, GNNQQNY that forms amyloid structures (PDB: 1YJP). The β -strands are oriented across the fiber axis. The peptides are in register (*top left*). Not only main chain hydrogen bonds (black) but also side chain hydrogen bonds (green) are formed. Pairs of the β -sheet stack in an antiparallel way to form a completely dry interface with the neighbor sheet (*top right*). *Bottom left*: A simplified illustration of the arrangement of the peptides. (Reprinted with permission from Sawaya *et al.* (2007) Atomic structures of amyloid cross- β spines reveal varied steric zippers. *Nature* **447**: 453–457). *Bottom right*: Between the pairs of β -sheets a wet interface is formed (water molecules are shown as red dots in lower figure).

3.3.3.1 Transient and long-term interactions

Oligomeric proteins are usually stable molecular aggregates with a long lifetime. An oligomeric enzyme may not need the interaction between the subunits for their function but rather to gain stability. On the other hand, for some oligomeric enzymes allosteric interactions between the subunits regulate the function of the protein (see Chapter 8). Stable oligomers, for example, virus coats or filamentous actin, can also have a dynamic nature and can polymerize from monomers and depolymerize into monomers to perform certain functions.

Many proteins are involved in short-term interactions. Enzymes belong to this class of proteins (see Chapter 8). They bind their substrates, in some cases other macromolecules, perform their enzymatic activity and dissociate. In the process the substrate or the enzyme may change its conformation, which may lead to a decreased affinity between the molecules. Many proteins without enzymatic activity also engage in temporary interactions with other macromolecules. For instance, adaptor proteins in signaling pathways bring together other proteins to allow them to use their enzymatic activity on the desired substrate (see Chapter 14).

3.3.4 Coenzymes and Metal Ions

3.3.4.1 Prosthetic groups and coenzymes

Many proteins, in particular enzymes, can only exert their biochemical function if they bind a non-protein molecule first. Such cofactors can be inorganic or organic. In some cases they are strongly bound to the protein such as metal ions like zinc, copper or heme groups. They are then called prosthetic groups. Other groups react with the substrates and need to be released for the next reaction. They are normally called *coenzymes* but also sometimes cosubstrates. Some examples are ATP, nicotinamide adenine dinucleotide (NAD⁺) and S-adenosyl methionine.

3.3.4.2 Metal binding

Metals are essential components of many biological systems, proteins as well as nucleic acids. About 40% of proteins in Protein Data Bank contain metals. Table 3.7 gives a summary of the roles and frequent protein ligands of the different metals. Metals can stabilize three-dimensional structures by cross-linking different parts of the structure; sometimes neutralizing negative charges (such as Asp or Glu residues) that otherwise would repel each other. Many enzymes frequently use transition metals for their catalytic activities, since these metals can readily cycle between different oxidation states. The metals that are bound as cations to proteins are: Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu and Zn. The side chains of proteins that bind metals are: Asp, Asn, Glu, Gln, Ser, Thr, His, Cys and more rarely Tyr and Met. The amino or carboxyl ends of a polypeptide may be engaged in metal binding, and main chain carbonyl oxygens often act as metal ligands for Na⁺, Mg²⁺, K⁺, Ca²⁺ and Mn²⁺. Water molecules are frequently ligands to metal ions and sometimes they can be deprotonated for participating in a hydration reaction as in carbonic anhydrase (see Section 8.1). The metals always bind to lone pairs of the ligands. Sometimes Asp, Glu and His can engage in the binding of two metal ions through two of the side chain atoms. Asp and Glu can also form bidentate interactions to a single metal atom.

TABLE 3.7 Metals in Proteins: States, Ligands and Roles

| Metal | Normal Oxidation States | Number of Ligands | Frequent Protein Ligands | Distance Metal- Ligand (Å) | Main Biological Roles | Figure Showing Example |
|-------|-------------------------------|-------------------------|---|----------------------------------|--|------------------------------|
| Na | I | 6 | C=O, H ₂ O | 2.4 | Charge carrier; osmotic balance | 8.24 |
| K | I | 8 | C=O, H ₂ O | 2.8 | Charge carrier; osmotic balance | 13.5 |
| Mg | II | 6 | Asp, Glu, Ser, Thr, C=O, H ₂ O | 2.1 | Structure stabilization, hydrolase, isomerase | 5.51, 16.7 |
| Ca | II | 6-7 | Asp, Glu, Ser, Thr, C=O, H ₂ O | 2.4 | Structure stabilization, trigger, charge carrier | 15.9, 15.16, 16.7 |
| Mn | II, III | 5-6 | Asp, Glu, His, H ₂ O | 2.2 | Photosynthesis integrins | 16.7 |
| Fe | II, III | 5-6 | His, Cys, Asp, Glu, Tyr, H ₂ O | 2.1 | Structure stabilization, oxidase, dioxygen transport and stor- age, electron trans- fer, nitrogen fixation radical generation | 3.17 |
| Ni | II | 4-6 | His, Cys, H ₂ O | 2.3 | Hydrogenase, hydrolase | 3.17 |
| Cu | I, II | 3-4 | His, Cys, Met, Asp | 2.0 | Oxidase, dioxygen transport, electron transfer | |
| Zn | II | 4-6 | His, Cys, Asp, Glu, H ₂ O | 2.0 | Structure stabilization, hydrolase | 8.4, 10.15, 10.16 |

Different metals prefer different ligands from the proteins and even the same metal can prefer different ligands depending on its oxidation state, so the binding strength can differ significantly. Thus, zinc has very strong binding to the sulfur atom of cysteines (Figure 3.17). It also binds strongly to the nitrogen atoms of histidine residues. Calcium, on the other hand, never binds to sulfur and rarely to nitrogen. It binds to oxygen atoms, but more weakly. Since calcium ions can bind and be released, it can function as a signaling molecule.

The coordination of different metals is also quite variable. Magnesium is a distinct case, which requires an octahedral environment of oxygens at short distances. Other metals, such as potassium, have more variability in their coordination geometry. The bond lengths between different metals and their ligands also vary.

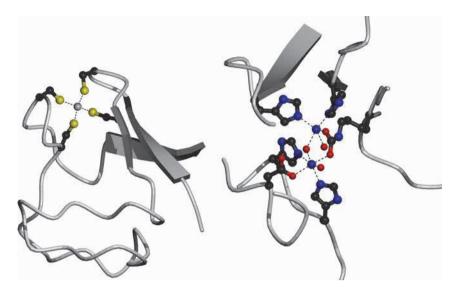


Fig. 3.17 ■ Metal binding. *Left*: Rubredoxin (PDB: 7RXN). Four Cys residues bind to a single iron ion. Right: Urease (PDB:2UBP). Two nickel ions are each bound by two nitrogen atoms from histidine residues. They both bind to oxygen atoms from a single carboxylysine residue. One of them binds to an oxygen atom from an aspartate, and they both bind to two solvent molecules (one shared).

3.3.4.3 Inorganic or organic metal clusters

Some proteins, in particular enzymes, sometimes contain clusters that include several metal atoms or ions. One class of these is the FeS clusters bound by cysteinyl residues, which often feature in electron transfer proteins. There are cases with two, three and four iron ions where inorganic sulfur atoms act as bridging ligands and Cys residues bind the irons to the protein. These complexes are 2Fe-2S, 3Fe-4S and 4Fe-4S. In the last two of these, the Fe and S atoms are located in the corners of a cube (Figure 3.18). When there are only three iron atoms one corner is missing.

One extreme case is ferritin, which is an iron storage protein. The protein shell has 24 subunits of the four-helical bundle type of fold arranged with 432 symmetry. The inner cavity has a diameter of 75Å and can accommodate ferric oxide with more than 4000 iron atoms. In addition, there is a certain amount of phosphate in the core. The inner side of the protein shell is also generally hydrophilic.

A range of organic complexes with metals is also part of proteins and in particular enzymes. Prime examples are the porphyrins that can bind metals to become heme with iron, or chlorophyll with bound magnesium.

Metals provide unique means to participate in oxidation/reduction reactions. Metals with several accessible oxidation states participate in such reactions. Iron, manganese and copper are of particular interest in this respect.

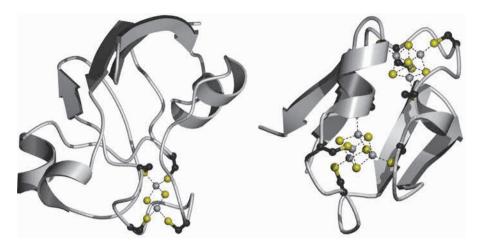


Fig. 3.18 ■ Metal binding in ferredoxin (2Fe-2S, PDB: 1FRR) and ferredoxin 2 (4Fe-4S, PDB: 1DUR).

For Further Reading

Original Articles

Bourgeois D, Vallone B, Arcovito A, *et al.* (2006) Extended subnanosecond structural dynamics of myoglobin revealed by Laue crystallography. *Proc Natl Acad Sci USA* **103**: 4924–4929.

Eriksson AE, Baase WA, Zhang X-J, et al. (1992) Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*: **255**: 178–183.

Gassner NC, Baase WA, Mooers BHM, *et al.* (2003) Multiple methionine substitutions are tolerated in T4 lysozyme and have coupled effects on folding and stability. *Biophys Chem* **100**: 325–340.

Lo Conte L, Chothia C, Janin J. (1999) The atomic structure of protein-protein recognition sites. *J Mol Biol* **285**: 2177–2198.

Prangé T, Schiltz M, Pernot L, *et al.* (1998) Exploring sites in proteins with xenon or krypton. *Prot Struct Funct Genet* **30**: 61–73.

Reviews

Branden C, Tooze J. (1999) *Introduction to Protein Structure*, 2nd edn. Garland Publishing, New York. Chothia, C. (1984) Principles that determine the structure of proteins. *Annu Rev Biochem* **53**: 537–572. Eisenberg D, Jucker M. (2012). The amyloid state of proteins in human disease. *Cell* **148**: 1188–1203. Harding MM, Nowicki MW, Walkinshaw MD. (2010). Metals in protein structures, a review of their principal features. *Cryst Rev* **16**: 247–302.

Marsh JA, Teichmann SA. (2015) Structure, dynamics, assembly and evolution of protein complexes. Ann Rev Biochem 84: 551-575.

Oldfield CJ, Dunker AK. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. Ann Rev Biochem 83: 553-584.

Razvi A, Scholtz JM. (2006) Lessons in stability from thermophilic proteins. Prot Sci 15: 156–1578. Richardson JS. (1981) The anatomy and taxonomy of protein structure. Adv Prot Chem 34: 167–339.

Basics of Membrane Proteins

4.1 Introduction

Both prokaryotic and eukaryotic cells are bounded by biomembranes that separate and protect cells from their environments. Eukaryotic cells also possess internal membranes that create discrete intracellular compartments, allowing the spatial separation of different chemical processes and intracellular resources. Embedded within biomembranes are a host of resident proteins that perform many essential functions. For example, they allow cells to acquire nutrients and expel waste; they sense and respond to external stimuli; they maintain and selectively dissipate electrochemical gradients; and they permit communication between different cells in multicellular organisms. Membrane proteins are also virulence factors in many pathogenic bacteria, where they facilitate adhesion onto host cells, confer antibiotic resistance, and mediate the uptake of trace metals that are required for survival.

Because of the importance of membrane proteins in so many diverse biological processes, they are the targets of many therapeutic antibodies and around half of today's small-molecule drugs — a fact that makes these proteins particularly attractive subjects for structural studies. Unfortunately, however, the structures of membrane proteins are notoriously difficult to determine. Indeed, less than 1.5% of the structures in the Protein Data Bank are membrane proteins, yet they represent approximately 30% of all genes in all organisms. What are the reasons for this disparity?

Firstly, membrane proteins are typically expressed at much lower levels than their soluble counterparts, making it difficult to obtain the quantities necessary for structural studies. Secondly, membrane proteins must be extracted out of their native lipid bilayer (Chapter 6) prior to purification and crystallization, and this requires the use of oftenharsh detergents that can adversely affect protein stability. Finally, growing membrane protein crystals that are of sufficient quality for X-ray diffraction studies provides its own unique set of hurdles, and a typical membrane protein structure can take years to

determine. Despite these challenges, the biological importance of membrane proteins has encouraged an ever-increasing number of ambitious researchers to pursue their structures, and today we are seeing the fruits of their efforts. This chapter will summarize our current knowledge on the characteristic structural features of this important protein family.

4.2 The Amphipathic Membrane Environment

Biological membranes are composed of amphipathic lipid molecules (Chapter 6), and therefore the bilayers themselves are also amphipathic in character (Figure 4.1). This feature provides a unique matrix for accommodating resident membrane proteins, which will encounter two polar regions arising from the lipid head groups as well as the hydrophobic environment of the lipid fatty acid tails. The hydrophobic membrane interior is approximately 30 Å thick, and it is highly unfavorable for polar or charged species to exist within this region (Figure 4.1). As we shall see in Section 4.4.1, this simple fact imposes some strict constraints on the secondary structures of all membrane-spanning proteins.

The two polar portions of the membrane are called the *interfacial regions*, since they exist at the interface between the hydrophobic membrane core and the extra-membranous bulk water. Contrary to what one might imagine, there are no discrete boundaries

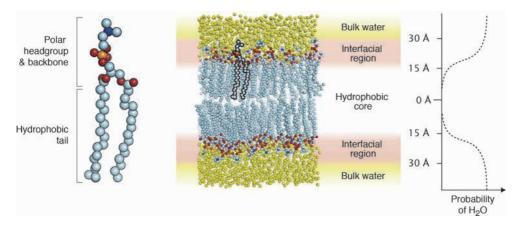


Fig. 4.1 ■ The amphipathic membrane environment. *Left:* The lipid phosphatidylcholine (PC) (Chapter 6). Hydrogen atoms have been omitted for clarity. Carbon atoms are colored light blue, nitrogens are dark blue, and phosphorus and oxygen atoms are orange and red, respectively. *Right:* A snapshot of the dynamic lipid bilayer. Water molecules are colored yellow. The probability curve shows that water does not reside permanently within the membrane core. However, it should be noted that water molecules can still pass transiently through the membrane during osmosis. The bilayer coordinates are reported in Wennberg CL, van der Spoel D, Jochen S. (2012) Large influence of cholesterol on solute partitioning into lipid membranes. *J Am Chem Soc* 134: 5351–5361.

between the membrane core and the interfacial regions, or between the interfacial regions and the bulk water. Instead, the interfacial regions contain a dynamic mixture of water, ions, and lipid backbones, fatty-acid chains, and head groups. These heterogeneous regions cover approximately 15 Å on either side of the hydrophobic membrane interior (Figure 4.1), and their polarity increases with increasing distance from the hydrocarbon core.

The amphipathic environment of cellular membranes is therefore vastly different from the aqueous milieu of the cytosol in which globular proteins fold and function. It follows that proteins traversing the membrane have evolved some distinctive structural characteristics suitable for interaction with both its hydrophobic interior and polar interfacial regions, and these properties will be explored in Sections 4.6 and 4.7.

The First Membrane Protein Structures

The first major structural insight into a membrane protein came in 1975, when Henderson and Unwin presented a seminal electron microscopy study on two-dimensional crystals of the protein bacteriorhodopsin embedded in its native bilayer ("the purple membrane"). From a series of electron diffraction patterns they derived a three-dimensional electron density map of the protein at 7Å resolution, which showed seven cylindrical rods of

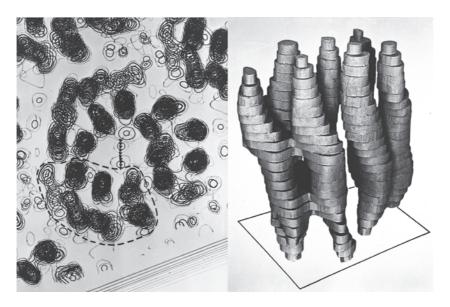


Fig. 4.2 ■ The first structural insights into a membrane protein. Left: Projection map of bacteriorhodopsin obtained through electron microscopy. The dashed line outlines a single copy of the protein. Right: 3D reconstruction of bacteriorhodopsin viewed from within the membrane plane. (Reprinted with permission from Henderson R, Unwin PNT. (1975) Three-dimensional model of purple membrane obtained by electron microscopy. Nature 257:28–32. Copyright (1975) Nature.)

density traversing the membrane, each about 35–40 Å long (Figure 4.2). These, it was concluded, represented seven closely packed transmembrane α -helices, giving us our first glimpse into how the α -helical class of membrane proteins (Section 4.6) are embedded within the bilayer.

Encouraged by these results, researchers dedicated their energies towards obtaining a higher-resolution structure in which the molecular details of the polypeptide could be discerned. However, this would first require growing three-dimensional crystals of purified protein that had been extracted out of the membrane. This was achieved by Michel and Osterhelt in 1980, who removed bacteriorhodopsin from its native bilayer by treatment with detergent molecules — the method still used by structural biologists today. Virtually simultaneously, the crystallization of another membrane protein using the same detergent extraction method was also reported, this time with $E.\ coli\ OmpF$ from the β -barrel class of proteins (Section 4.7). Unfortunately, however, the crystals of these two proteins still did not readily yield high-resolution structures.

The turning point came in 1985, when, after switching to the study of the photosynthetic reaction center from *Rhodopseudomonas viridis*, Michel and his coworkers reported the first ever atomic-level structure of a membrane protein. The structure of the reaction center with all its ligands and metal cofactors revealed a new world of electron chains, transport pathways across the membrane, and interactions with lipid molecules — a phenomenal achievement even by today's standards — and Michel shared the 1988 Nobel Prize in Chemistry with his colleagues Deisenhofer and Huber.

Finally, in 1990, a high-resolution structure of a β -barrel protein was also determined, thereby completing our basic structural understanding of the two classes of proteins that reside within the membrane.

4.4 Structural Classification of Membrane Proteins

4.4.1 Peripheral and Integral Membrane Proteins

Since the first pioneering studies on the structures of membrane proteins, our understanding of these remarkable macromolecules has increased considerably. Membrane proteins are now classified into a number of different families according to their secondary structures and how they associate with the bilayer.

At the broadest level, membrane proteins are classified as either *peripheral* or *integral*, and two different working definitions for each of these terms are currently in use. In the first definition, and that which is used throughout this text, peripheral membrane proteins are associated with only one side of the bilayer, whereas integral membrane proteins fully span its width (Figure 4.3). The second definition is based on a protein's behavior when membranes are treated with high salt buffer (0.5–1.0 M). Here, peripheral

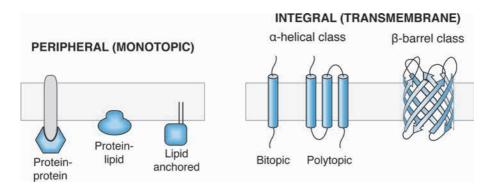


Fig. 4.3 \blacksquare Types of membrane proteins. The α -helical integral membrane proteins can be further divided into the additional subfamilies described in Section 4.4.2, Figure 4.4.

membrane proteins are those that dissociate from the bilayer in high salt, while integral membrane proteins remain bound to the bilayer (the molecular basis of this behavior is described in Section 4.5.1). It is important to note that the two classification systems are not always synonymous, since some proteins that do not traverse the bilayer can still be resistant to removal by high salt if their structures penetrate significantly into the hydrophobic core of the membrane.

Peripheral membrane proteins (also called *monotopic* membrane proteins) can reside permanently at the bilayer or can be temporarily recruited in response to some signaling event. This latter sub-group is known as conditional membrane proteins. Peripheral membrane proteins can localize to the bilayer in three different ways (Figure 4.3): (1) through interactions with integral (membrane-spanning) membrane proteins, (2) through specific or non-specific interactions with membrane lipids, or (3) through covalently attached lipid chains or glycophosphatidylinositol (GPI) moieties that become embedded in the membrane. Many peripheral membrane proteins use more than one of these methods to associate with the bilayer.

Integral membrane proteins are also known as transmembrane proteins, since they fully span the bilayer. Remarkably, all transmembrane proteins fall into one of only two structural classes — the α -helical and the β -barrel proteins — which are distinguished by the secondary structures of their membrane-spanning regions (Figure 4.3). α -Helical transmembrane proteins can span the bilayer once (bitopic or single-pass) or multiple times (polytopic or multiple-pass). They can also be further divided into the subfamilies described in Section 4.4.2. In contrast, there are no additional structural subfamilies in the β-barrel class, and the backbones of these proteins always cross the membrane multiple times (Section 4.7 and Figure 4.3).

Considering that most soluble proteins are comprised of a diverse mixture of helices, sheets and coils (Section 2.3), it might at first seem surprising that integral membrane proteins are only ever built upon one of two mutually exclusive structural frameworks. However, this is soon explained when we recall that the membrane core is a hydrophobic environment that disfavors the presence of polar species like an exposed polypeptide

backbone (Section 4.2). Peptide bonds have significant hydrogen-bonding potential, and it is energetically costly to remove these groups from the aqueous cytosol and partition them into the hydrophobic membrane interior. The α -helical and β -sheet secondary structures circumvent this problem, since all polar backbone groups are involved in intramolecular hydrogen bonds that fully-satisfy their hydrogen bonding-potential (Sections 2.3.1.1 and 2.3.2.1).

4.4.2 Types of Helical Integral Membrane Proteins

The α-helical membrane-spanning proteins can be classified beyond simply "bitopic" or "polytopic". Most commonly, bitopic membrane proteins are further divided into *Type I* or *Type II* subgroups, while all polytopic proteins would be designated *Type III* (Figure 4.4) (although in practice we rarely refer to polytopic membrane proteins in this way). Types I and II bitopic membrane proteins differ in their orientations within the membrane — the N-terminus is extracytosolic in Type I and cytosolic in Type II.

Bitopic membrane proteins can be further classed as either stop-transfer anchored, signal-anchored, or tail-anchored (TA) (Figure 4.4). *Stop-transfer anchored* proteins possess a signal peptide at their N-terminus that is responsible for membrane targeting and that is removed in the mature protein. The transmembrane domains from these proteins perform two different functions: they stop the transfer of the polypeptide across the

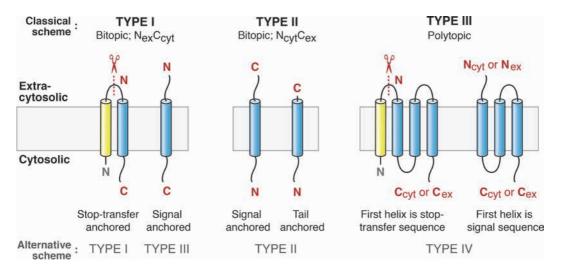


Fig. 4.4 ■ Classification of helical membrane proteins. The most common classification scheme is shown above the figure in bold type. An alternative scheme (not addressed in this book but sometimes used elsewhere) is shown below the figure. Signal peptides are represented by yellow helices and these are cleaved off in the mature protein. For polytopic proteins, " N_{cyt} or N_{ex} " indicates that the N-terminus can lie *either* inside or outside of the cytosol (and the figure represents only one of these scenarios). The same nomenclature is used for the C-terminus.

membrane during translation (Section 4.6.7.2), and they anchor the mature protein to the membrane. Hence they are called "stop-transfer anchored" proteins, and their transmembrane helix is often referred to as a *stop-transfer sequence*. Mature stop-transfer anchored proteins will always have their N-terminus outside of the cytosol in a Type I orientation (more details in Section 4.6.7.2). Polytopic membrane proteins can also possess signal peptides, and their first transmembrane helix acts as a stop-transfer sequence. It follows that the polytopic membrane proteins with a signal peptide will also always have their N-terminus outside of the cytosol in the mature protein (Figure 4.4).

In contrast, *signal-anchored* bitopic proteins do not possess a signal peptide. Instead, they are targeted to the membrane through their transmembrane domain that itself acts as an internal signal sequence (also called a signal-anchor sequence). Such proteins can adopt either a Type I or a Type II orientation (Figure 4.4) depending on their sequence and structural characteristics (Section 4.6.6.1). Most polytopic membrane proteins are targeted to the bilayer via internal signal-anchor sequences (i.e. their first transmembrane helix), rather than signal peptides.

Finally, tail-anchored (TA) membrane proteins are Type II bitopic proteins that possess a large N-terminal domain and only a very short C-terminal tail. Usually, their membrane-spanning domains are located within no more than ~30 amino acids of the C-terminus. Tail-anchored bitopic proteins are unique because they are inserted into the membrane in a different manner than their signal-anchored and stop-transfer anchored counterparts (Section 4.6.7.1).

Peripheral Membrane Proteins

4.5.1 The Importance of Electrostatics

Peripheral membrane proteins can bind to the bilayer in a variety of ways. Some interact with the lipid head groups only, while others bury a portion of their structure into the hydrophobic core of the membrane. Some bind specifically to certain lipids, while others non-specifically recognize the broader chemical or physical properties of the wider membrane. Yet despite the diversity of these interactions, the structures of peripheral membrane proteins have revealed some common themes in how they recognize and associate with biomembranes.

One typical feature of peripheral membrane proteins is that their membrane-binding faces are often rich in the basic (positively charged) residues arginine and lysine (Figure 4.5). This is because membrane surfaces generally display a net negative charge, since biomembranes are comprised of a combination of uncharged (e.g. glycolipids), zwitterionic (e.g. PC and PE), or anionic lipids (e.g. PG, PS and PI) (Section 6.1.1). It follows that when positively charged residues are clustered on one face of a peripheral membrane

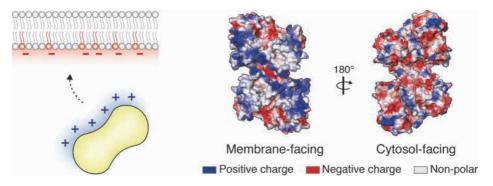


Fig. 4.5 ■ Electrostatic interactions between peripheral membrane proteins and biomembranes. *Left:* The positively charged faces of many peripheral membrane proteins are attracted to negatively charged membrane surfaces. *Right:* The electrostatic surface potential of the peripheral membrane protein squalene-hopene cyclase (PDB: 2SQC). Positively charged residues on its membrane-interacting face contact the anionic lipid head groups and non-polar regions are buried in the hydrophobic membrane interior.

protein along with few anionic residues, a favorable electrostatic interaction can be formed with negatively charged membranes (Figure 4.5). This non-specific electrostatic attraction can also help to recruit conditional membrane proteins (Section 4.4.1) to the membrane surface, where they can then (for example) initiate specific interactions with particular lipid head groups.

Those peripheral membrane proteins that rely primarily on electrostatic interactions for membrane binding can usually be dissociated in the laboratory by treatment with high salt buffer, which disrupts charge-charge contacts. This is in contrast to peripheral membrane proteins that have extensive hydrophobic interactions with the lipid fatty acid chains in the membrane interior. Proteins such as these can only be removed from the bilayer by treatment with detergent molecules that mimic the amphipathic membrane environment.

4.5.2 Amphipathic Helices in Membrane Binding

4.5.2.1 The roles of amphipathic helices

Amphipathic helices possess a spatially biased distribution of polar and non-polar residues, giving rise to two opposing faces with markedly different properties. In Section 3.3.2, we saw how amphipathic helices in soluble proteins can be exploited to form extended coiled-coil structures. In peripheral membrane proteins, however, amphipathic helices play an entirely different role. When positioned parallel to the membrane surface in the interfacial region (Section 4.2), their polar faces can interact

with the lipid backbones and head groups while their non-polar faces contact the hydrophobic fatty acid tails (Figure 4.6). These helices therefore provide the perfect structural scaffold for membrane binding and are frequently observed in the structures of peripheral membrane proteins.

In some peripheral membrane proteins, amphipathic helices can play a special role in sensing membrane curvature. Depending on their particular sequence characteristics, some amphipathic helices will bind preferentially to highly curved membranes such as vesicles. This is because high membrane curvature decreases the lipid packing density on the outer leaflet of the vesicle and introduces packing defects, allowing helices to insert more readily. In this way, some peripheral membrane proteins are recruited to the correct membrane on the basis of curvature (see also Section 6.3.5, on BAR proteins).

Amphipathic helices can also sometimes play a role in *generating* membrane curvature, since they are inserted into only one of the two bilayer leaflets. This can cause the expansion of one leaflet relative to the opposing leaflet, thereby mechanically inducing a higher curvature (Figure 4.7). One example of this is the effect of the enzyme CTP: phosphocholine on the bilayer, discussed in Section 6.3.4.1.

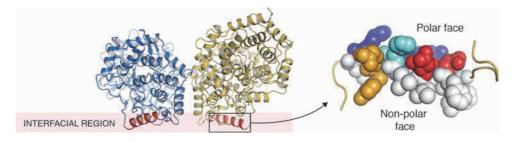


Fig. 4.6 ■ Amphipathic helices in peripheral membrane proteins. Left: The structure of the homodimeric peripheral membrane protein squalene-hopene cyclase (PDB: 2SQC). Its amphipathic helices are colored red. Right: Close-up of the amphipathic helix from squalene-hopene cyclase shown as spheres. Non-polar amino acids are shown in white, polar residues are orange, acidic residues are red, and basic residues are blue (the weakly-basic histidine is in light blue).

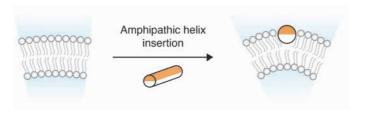


Fig. 4.7 ■ Amphipathic helices can influence membrane curvature. In the right-hand panel, the helix is viewed down its long axis. Its polar face is colored orange and its hydrophobic face is white.

4.5.2.2 Identifying amphipathic helices

When their sequences are written linearly, the amphipathic helices of peripheral membrane proteins are much more difficult to recognize than the simple heptad repeats of soluble coiled-coils (Section 3.3.2). This is because membrane-binding amphipathic helices do not display a recurring sequence pattern. However, they still exhibit an overall spatial bias of polar and non-polar residues when arranged in a three-dimensional helix (Figures 4.6 and 4.8). Fortunately, amphipathic helices can be easily identified in *helical wheel diagrams*, which show the distribution of side chains when a helix is viewed down its axis. Helical wheels are constructed by mapping the amino acid sequence of a helix around the circumference of a circle, with each successive residue in the sequence separated by 100° (since there are 3.6 residues per full 360° turn of an α -helix; Section 2.3.1.1). The relationship between primary structure, α -helix geometry and the helical wheel is illustrated in Figure 4.8.

When using helical wheel diagrams to predict amphipathic helices, the degree of amphipathicity is quantified by the *helical hydrophobic moment*. The hydrophobic moment indicates the magnitude of spatial bias in the distribution of polar and non-polar amino acids about the helix axis.

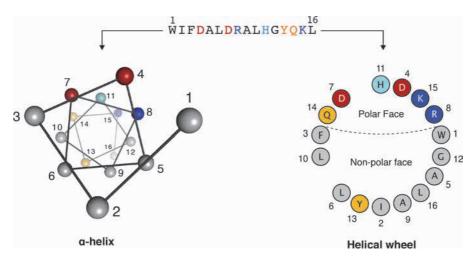


Fig. 4.8 ■ Identifying amphipathic helices through helical wheel diagrams. *Top*: The sequence of the membrane-interacting amphipathic α -helix from squalene-hopene cyclase, whose structure is shown in Figure 4.6. *Left*: The position of each residue when arranged in an α -helix. The helix is viewed down its axis with $C\alpha$ atoms shown as spheres. *Right*: The same sequence represented as a helical wheel. For longer helices, the wheel would continue with consecutive residues placed 100° apart. Amino acids are colored as for Figure 4.6.

4.6 α-Helical Integral Membrane Proteins

4.6.1 Occurrence and Structural Diversity

The helical class of transmembrane proteins (α -TMPs) exploits α -helical secondary structure to traverse the bilayer. These proteins are approximately 10 times more common than β -barrel integral membrane proteins (Section 4.7), and represent ~20–30% of all genes from all organisms. Almost all membranes from both prokaryotes and eukaryotes contain a wide variety of helical membrane proteins. However, only a few known examples exist in the outer membranes of gram-negative bacteria (where the β-barrels instead predominate).

Helical membrane proteins display great diversity in their structures and some examples are shown in Figure 4.9. As we saw in Section 4.4.1, α-TMPs can be either *bitopic* and cross the membrane only once, or they can be *polytopic* and span the membrane more than once. Polytopic α -TMPs can possess either an odd or an even number of transmembrane helices, though they usually have less than 14 in a single polypeptide chain. Some helical membrane proteins form higher-order homo- or hetero-oligomers, and others contain very large soluble domains (Figure 4.9). Today, recombinant protein expression methods often allow the structures of soluble domains to be determined separately from the transmembrane regions with their associated technical challenges.

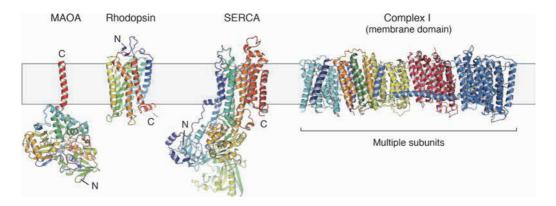


Fig. 4.9 \blacksquare Examples of α -helical membrane proteins. The structures shown are monoamine oxidase A (MAOA; PDB: 2Z5X), rhodopsin (1F88), the sarcoplasmic reticulum Ca²⁺–ATPase (SERCA, 3B9B), and the membrane domain from respiratory Complex I (4HE8). The membrane domain from Complex I is composed of seven individual subunits that are shown in different colors.

4.6.2 Features of Simple Transmembrane α-helices

The basic structural unit of all helical membrane proteins is the α -helix (Section 2.3.1.1). There are two fundamental properties of the α -helix that are of particular importance to the structures of α -TMPs: that all side chains point outwards from the helix axis, and that the rise per residue is 1.5 Å. These two features dictate which particular amino acids are favored in transmembrane helices and how many residues are necessary to span the membrane.

Because every side chain in an α -helix points outwards, every residue in the transmembrane region will exist in one of two possible environments: (i) exposed to the lipid fatty acid chains, or (ii) at a helix-helix interface (Figure 4.10, left panel). Hydrophobic residues predominate at *both* of these interface types, and therefore membrane-spanning helices consist of long continuous stretches of hydrophobic residues such as phenylalanine, leucine and isoleucine. The 1.5 Å rise per residue in an α -helix means that about 21 hydrophobic or non-polar residues are required to cross the ~30 Å fatty acid core of the bilayer. In practice, however, most transmembrane helices are tilted 10–30° relative to the bilayer normal, and their hydrophobic portions are therefore somewhat longer (usually

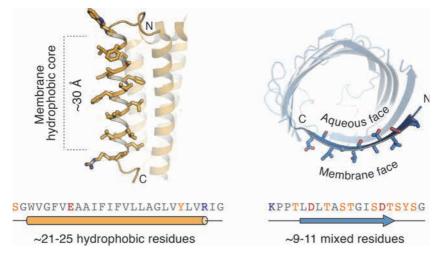


Fig. 4.10 Features of membrane-spanning α -helices and β -strands. *Left:* A transmembrane α -helix viewed from within the membrane plane. Residues on the left side of the helix contact the lipid fatty acid chains and residues on the right contact adjacent α -helices. Coordinates are from the structure of respiratory Complex I (PDB: 3RKO). *Right:* A β -strand from a β -barrel aqueous pore, viewed from above the membrane. The small sphere on the strand is representative of an inward-facing glycine residue (since glycine possesses only a hydrogen atom for a side chain). Some β -strands from the barrel have been omitted for clarity. Coordinates are from the protein TolC (1TQQ). The sequences of the helix and strand from each panel are shown below the figures, with amino acids colored as for Figure 4.6.

21–25 residues in length). Transmembrane helices less than 20 residues are rare but also exist, as do highly tilted or bent helices that are longer than 35 residues.

The long and continuous hydrophobic regions of transmembrane helices are easily recognizable in the amino acid sequences of α-TMPs, allowing membrane-spanning segments to be predicted with relative ease and accuracy. This is in stark contrast to the situation for the β -barrel membrane proteins. As we shall see in Section 4.7, β -barrels have much shorter transmembrane regions (only ~9-11 residues) and also demonstrate an alternating inside-outside arrangement of their side chains (Figure 4.10, right panel). As a result, their component β-strands possess an inconspicuous mixture of hydrophobic and polar residues that are difficult to distinguish from non-membrane-spanning segments.

Although non-polar and hydrophobic amino acids predominate in membranespanning helices, polar and charged residues can also sometimes occur within the membrane core. However, these are rarely directed towards the lipid tails but rather mediate helix-helix hydrogen bonds or salt bridges. In fact, the majority of transmembrane helical interfaces are stabilized by at least one interhelical hydrogen bond or salt bridge.

Polar residues in the transmembrane region can also point towards the interior of the protein and play an important functional role. For example, they can create binding sites for ions in transporters such as the P-type ATPases (Section 13.3.1.1), provide a pathway for polar solutes in channel proteins (Section 13.2.2), or mediate cofactor binding in integral membrane enzymes and electron transfer proteins.

4.6.3 Helix-Helix Interactions Within the Membrane

4.6.3.1 The geometry of helix-helix pairs

In order for a membrane protein to adopt its functional three-dimensional structure, its component helices must correctly associate with each other within the bilayer. But before examining the molecular details of helical interfaces, we will first consider how helix pairs can be classified by their overall geometries.

The transmembrane helices in polytopic or multimeric bitopic α-TMPs adopt a tertiary structure known as the *helical bundle*, in which all helices are oriented roughly parallel to each other (Figure 4.9). However, because of side chain packing constraints, almost all transmembrane helices cross each other at a small angle. This gives helical interfaces an inherent handedness. The simplest means of determining the handedness of an interface is to view it from within the plane of the membrane (Figure 4.11). If the front helix is tilted to the left relative to the rear helix, the interface is left-handed. If it is tilted to the right, then it is right-handed. Because the two interacting helices can also lie either parallel or antiparallel to each other, this leads to four possible interface types: one each of left- and right-handed parallel, and also left- and right-handed antiparallel (Figure 4.11).

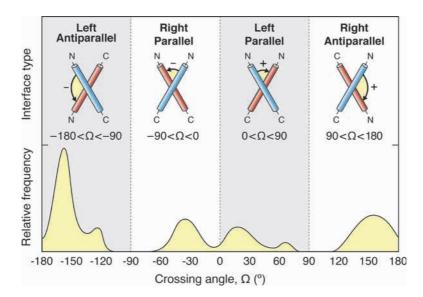


Fig. 4.11 ■ Helical interface types and their approximate frequencies. The helical crossing angle (Ω) is measured from N-to-N (or C-to-C) termini. The approximate frequencies of different crossing angles are from the Transmembrane Protein Helix-Packing Database (http://biocluster.iis. sinica.edu.tw/TMPad), in which the crossing angles of all interfaces in all known membrane protein structures (until 2011) have been calculated.

Into which of the four classes an interface falls is described by the helical crossing *angle* (or the *interhelical angle*; Ω). This is the signed angle between the two helix axes at their point of closest approach. Two conventions are used when reporting helical crossing angles. In the first, and that which is used throughout this book, the crossing angle lies between -180° and +180° and is the angle between the two like termini of each helix (either N-to-N or C-to-C; Figure 4.11). Obtuse angles therefore correspond to antiparallel helices, and acute angles to parallel helices. The angle has a positive value if, when measuring N-to-N or C-to-C, the rear helix is rotated clockwise relative to the front helix. Negative values indicate an anticlockwise rotation. Therefore, when the magnitude and sign of Ω are given, one can also deduce whether an interface is left- or right-handed.

In the second convention for reporting crossing angles, the directionality of the helices is ignored (that is, whether they are parallel or antiparallel) and the crossing angle is the acute angle between the two helices. This simplified crossing angle therefore lies between -90° and +90°, with the sign of the angle calculated as above. To fully describe the interaction, it must then be additionally stated whether the helices are parallel or antiparallel. For example, a helix interface with a crossing angle of –160° can also be expressed as "+20°, antiparallel".

As shown in Figure 4.11, the helical crossing angles in membrane proteins of known structure cluster into four major groups. This non-uniform distribution of crossing angles reflects the constraints of packing opposing side chains against each other whilst still

keeping both helices approximately parallel to the membrane normal. It is also evident from Figure 4.11 that the four main interface types do not occur with equal frequency. Instead, the most common helix packing geometry is left-handed antiparallel with a crossing angle of approximately -160° . This geometry — and its parallel equivalent with $\Omega \approx$ +20° — is the same as that observed in the coiled-coils of fibrous proteins. This is because membrane protein side chains pack in the same "knobs-into-holes" manner that is characteristic of coiled-coils, which is described in detail in Section 3.3.2.3 and Figure 3.13.

4.6.3.2 Sequence features of helix-helix interfaces

Even with the progress of recent years, we still lack a detailed understanding of the motifs and energetic forces that govern helix-helix association in membrane proteins. Nonetheless, it is clear that glycine, alanine and serine residues are overrepresented at many helical interfaces. The prevalence of glycine in transmembrane helices might at first seem surprising, since its high backbone flexibility and role as a "helix-breaker" means that it is virtually absent from the helices of soluble proteins.

So what do glycine, alanine and serine residues have in common that explains their abundance at transmembrane helical interfaces? Figure 2.2 shows that these residues possess small side chains, which in turn permits transmembrane helices to interact more closely than would be possible if large amino acids lay at the interface. The close association of membrane-spanning helices can be stabilizing in a number of ways. Firstly, it promotes favorable van der Waals contacts (Section 2.1.5) between opposing side chains. These contacts are an important driving force for helix oligomerization in membrane proteins, since their interfaces are dominated by interactions between non-polar side chains in a shape-complementary manner. The close association of two helices is also thought to allow the formation of networks of weakly stabilizing backbone-to-backbone Cα-H···O hydrogen bonds. These bonds are extremely rare in soluble proteins but can contribute to the stability of certain transmembrane helical interfaces. Finally, the hydroxyl group from interfacing serine side chains can hydrogen bond with other polar amino acids on an adjacent helix or with opposing backbone carbonyl groups.

In addition to these general sequence trends, one specific motif has now been identified that — along with other residues with appropriate shape complementarity or hydrogen-bonding potential — is known to mediate some helix-helix interactions in α -TMPs. This is the GxxxG (or Gx_3G) motif, which consists of two glycine residues that lie at the helix-helix interface, separated by any three amino acids. This sequence pattern has been extended to include the amino acids alanine and serine in place of glycine, so it is sometimes referred to as the Sm-xxx-Sm motif (where "Sm" represents any of the small residues glycine, alanine or serine).

The GxxxG motif can exist at all four interface types in Figure 4.11 and is often only found in one chain of the helix pair. The molecular details of the canonical GxxxG dimerization motif are typified by the bitopic membrane protein glycophorin A (GpA) in which

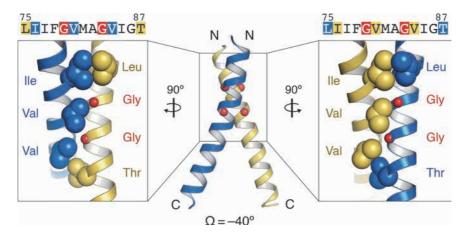


Fig. 4.12 • The GxxxG motif in glycophorin A. The $C\alpha$ atoms from the GxxxG glycines are shown as red spheres. The sequence of GpA about the dimerisation interface is shown. Coordinates are from the homodimeric membrane-spanning domain of human glycophorin A (PDB: 1AFO). Because the protein is a symmetric homodimer, the contacts exist twice across the interface.

the motif was first discovered. Because glycine lacks a side chain, the two glycine residues form a groove at the interface that can accommodate larger side chains from the opposing helix (Figure 4.12). In this way, close interhelical association is permitted.

4.6.4 Deviations from Simple α -helices

4.6.4.1 3_{10} and π -helices

Although the canonical α -helix is straight and undistorted (Figure 2.12), at least 50% of helical polytopic membrane proteins contain one or more of the non-canonical structural motifs illustrated in Figure 4.13. For example, 3_{10} and π -helices can sometimes occur within α -TMPs, leading to narrower and wider helices, respectively (Section 2.3.1.3). Like in soluble proteins, 3_{10} and π -helices in membrane proteins are usually shorter than 8 amino acids due to their diminished long-range stability. Because of this, there are no known examples of any one such helix traversing the entire bilayer. Rather, 3_{10} and π -helices occur as short stretches — usually no more than two helical turns — within the greater context of an α -type helix (and often following a helical kink; Section 4.6.4.2).

4.6.4.2 Helical kinks

Helical kinks are undoubtedly the most common structural irregularity in transmembrane helices. Kinked helices are those that exhibit a change in their axis direction at some specific point, accompanied by the local disruption of one or more backbone hydrogen

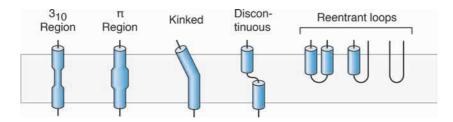


Fig. 4.13 ■ Non-canonical structural elements in transmembrane α -helices.

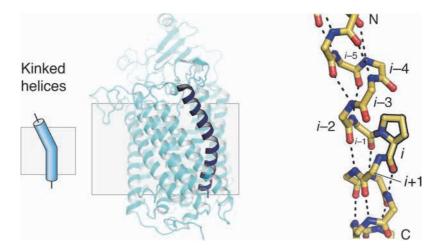


Fig. 4.14 The structural basis of proline-induced helical kinks. Coordinates are from the Paracoccus denitrificans aa₃-type cytochrome c oxidase (PDB: 1AR1; chain A, residues 218–251). The kink-inducing proline at position i is outlined in black. The kink angle is 34°. The i^{NH} to $i-4^{CO}$ and $i+1^{NH}$ to $i-3^{CO}$ hydrogen bonds have been broken (the carbonyl from residue i-3 is too far from residue i+1 to form a hydrogen bond). Note that although the proline residue at position i causes the kink, the kink itself is centred around residues i-4 and i-3.

bonds (Figure 4.14). Such kinks can be mild (<15°) or extreme (>50°), though typically fall into the range of 15-30°. Helical kinks fulfill two vital roles. Firstly, they increase membrane protein structural diversity by allowing transmembrane helices to adopt a noncanonical conformation. Secondly, the loss of backbone hydrogen bonds at the site of a helical kink can (in some cases) create a flexible hinge about which the protein can move during its transport or signaling cycle.

Around 35% of transmembrane helical kinks are caused by proline residues, which occur much more frequently in membrane-spanning helices than in the helices of soluble proteins (where proline is the least common amino acid in helices). The capacity of proline to kink a helix can be explained by its unique side chain — the cyclic pyrrolidine ring that incorporates the backbone nitrogen atom (Figures 2.2 and 4.14). This distinctive property induces a helical kink in two ways. Firstly, because the proline backbone lacks an amide hydrogen, it is unable to donate a hydrogen bond to the carbonyl group from the $(i-4)^{th}$ residue (numbered relative to the proline at position i; Figure 4.14). This disrupts the

canonical α -helical hydrogen bond network (Section 2.3.1.1) and gives the $(i-4)^{\text{th}}$ residue increased freedom. Secondly, the proline side chain is unable to project away from the helix axis due to its cyclic structure, causing a steric clash with the $(i-4)^{\text{th}}$ carbonyl group and further forces its reorientation. These structural deviations result in the local disruption of a minimum of one hydrogen bond: always the i^{NH} to $i-4^{\text{CO}}$ bond, usually also the $i+1^{\text{NH}}$ to $i-3^{\text{CO}}$ bond, and sometimes the $i-1^{\text{NH}}$ to $i-5^{\text{CO}}$ bond. Loss of these hydrogen bonds in turn permits kinking of the helix.

It should, however, be emphasized that although proline residues are the most common cause of helical kinks and generally produce the widest kink angles, kinks can also occur in the absence of proline. Furthermore, around one-third of prolines in transmembrane regions are accommodated *without* a kink, and the helix backbone displays only a minor distortion in the turn preceding the proline. Nonetheless, these non-kink prolines will still allow greater backbone flexibility and might still play a role in conformational changes and dynamics. It is therefore clear that other factors — such as tertiary helix-helix packing interactions — are important for kink generation in transmembrane helices.

4.6.4.3 Discontinuous helices and reentrant loops

Around 5% of all membrane-embedded residues do not possess α -helical secondary structure but instead form coils. This fact might at first seem surprising, given that it is highly energetically unfavorable for the polar polypeptide backbone to reside within the hydrophobic bilayer interior in an unfolded state (Section 4.2). However, the backbones of membrane-embedded coils are rarely in contact with the lipid fatty acid tails. Instead, they are usually buried within the protein core. In this way, they can border polar internal microenvironments or hydrogen bond with polar side chains.

Coil regions within the membrane can arise from either *discontinuous helices* or *reentrant loops*. Discontinuous helices — also known as *unwound helices* — contain a short stretch of coil resides within their membrane-embedded regions but still fully traverse the membrane (Figure 4.15). The coils of discontinuous helices are often functionally important. For example, in the Ca²⁺-ATPase and various secondary active transporters the coil backbone creates an internal binding site for transported ions (Section 13.3.1.1 and Figure 13.8). Discontinuous helices can also provide a favorable electrostatic environment for ion binding, since helical dipoles (Section 2.3.1.4) are exposed within the protein interior (Figure 4.15).

Helical membrane proteins can also possess *reentrant loops*, in which the polypeptide enters the membrane from one side, penetrates into the hydrophobic core, then reverses direction and exits the membrane on the same side (Figure 4.15). Reentrant loops come in three basic forms: helix-coil-helix, helix-coil (or coil-helix), and coil-only regions (Figure 4.13). Reentrant loops are generally less hydrophobic than membrane-spanning helices and can be difficult to detect from a protein's sequence. As for discontinuous helices, the coil regions or exposed helical dipoles of re-entrant loops often have

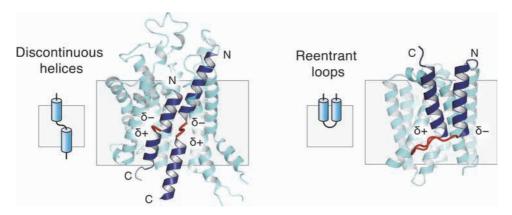


Fig. 4.15 ■ Discontinuous helices and reentrant loops. Left: Discontinuous helices in the sarcoplasmic reticulum Ca²⁺-ATPase (PDB: 3B9R). The membrane-embedded coil regions are colored red and helical dipoles are indicated. Soluble domains have been omitted for clarity. Right: A reentrant loop of the helix-coil-helix type from the uracil/proton symporter UraA (PDB: 3QE7).

functional significance in ion or substrate binding. One such example is in the protein aquaporin, which is described in Section 13.2.2.

4.6.5 The Interfacial Regions of Helical Membrane **Proteins**

4.6.5.1 Sequence features in the interfacial regions

In our exploration of the structures of helical membrane proteins thus far, we have focused our attention on the areas embedded within the hydrophobic core of the bilayer. However, membrane-spanning proteins must also pass through and interact with the polar interfacial regions of the bilayer (Section 4.2). Helical and coil secondary structures predominate in these regions, and the protein sequence is markedly different than in the membrane interior.

Polar residues are much more common in the interfacial region than in the hydrophobic core. This is because they can interact favorably with the lipid head groups and backbones. The amino acids tryptophan and tyrosine are particularly abundant, forming aromatic girdles (or aromatic belts) around each end of the protein (Figure 4.16). The prevalence of these residues can be explained by their unique side chains, which contain both aromatic rings and polar groups. The non-polar aromatic rings can interact with the lipid fatty acid tails, while the polar groups can hydrogen bond with the lipid head groups and backbones (Figure 4.11).

Some amino acids that are located near the (indistinct) boundary between the interfacial region and the hydrophobic core display some interesting orientational preferences.

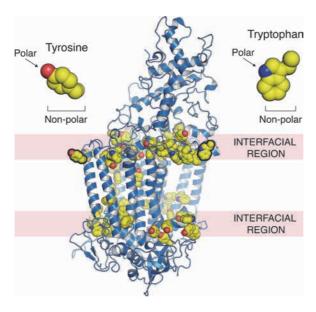


Fig. 4.16 ■ The aromatic girdles formed by tryptophan and tyrosine residues. Tryptophan and tyrosine side chains are shown as spheres. Coordinates are from the R. viridis photosynthetic reaction center (PDB: 1PRC).

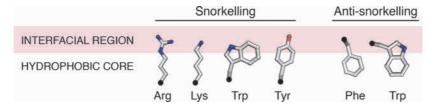


Fig. 4.17 ■ Common snorkeling and antisnorkeling preferences in some amino acids. The positions of $C\alpha$ atoms are indicated by black circles.

When arginine, lysine, tryptophan and tyrosine residues occur within but at the periphery of the hydrophobic membrane interior, their side chains are usually oriented such that their charged or polar groups are directed into the interfacial region (Figure 4.17). This phenomenon is known as *snorkeling*, and it allows the polar groups of these residues to interact with the lipid backbones and head groups while their non-polar portions remain within the hydrophobic membrane core. The inverse orientation is called antisnorkeling, which occurs when interfacially-located hydrophobic residues such as phenylalanine point their side chains into the aliphatic membrane interior (Figure 4.17). Because of their amphipathic side chains, tryptophan (and to a lesser extent tyrosine) residues typically snorkel when located within the membrane core but anti-snorkel when located within the interfacial region (Figure 4.17).

4.6.5.2 Interfacing helices

One striking structural element sometimes found in helical membrane proteins is the interfacing helix, which lies along the plane of the membrane in the interfacial region (Figure 4.18). These helices can exist at a protein's N- or C-termini, or they can connect two transmembrane helices. Interfacing helices are akin to the membrane-binding helices characteristic of many peripheral membrane proteins, as they frequently contain a spatially biased amphipathic mixture of amino acids (Section 4.5.2). In this way, they can interact with the hydrophobic membrane interior on one face, and the polar interfacial region on the other face (Figures 4.6 and 4.18).

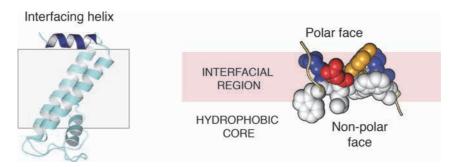


Fig. 4.18 ■ Interfacing helices in helical membrane proteins. *Left:* An interfacing helix (dark blue) from the inwardly rectifying potassium channel KirBac1.1 (PDB: 1P7B). Right: The same interfacing helix shown as spheres, demonstrating its amphipathicity and approximate location in the membrane. Residues are colored as for Figure 4.6.

4.6.6 The Topology of Helical Membrane Proteins

4.6.6.1 Determinants of topology

Having examined the individual elements that comprise a helical membrane protein its transmembrane helices, their interfaces and the interfacial structures — we can now consider how these elements are arranged in the full-length protein. The structure of a membrane protein is described at the simplest level by its *topology*, which is the number of its transmembrane segments and their directionality within the bilayer. The locations of the loop regions and the N- and C-termini (that is, whether they are cytosolic or extracytosolic) are also specified by a protein's topology. All of these features can be summarized in topology diagrams, which indicate the path of the protein backbone relative to the membrane (e.g. Figure 4.19, lower panels).

One of the major factors dictating the overall orientation of a helical membrane protein is the *positive-inside rule*, which states that the cytosolic ("inside") face will be that which possesses the greatest number of the positively charged residues arginine and lysine. This pattern holds true for both bitopic and polytopic α -TMPs across all organisms, but it is most pronounced in bacteria. The α -TMPs from the eukaryotic mitochondrial inner membrane and chloroplast thylakoid membranes also follow the positive-inside rule, with their positive faces then directed towards the insides of these organelles. The molecular basis of the positive-inside rule is not yet fully understood, but it appears to stem from interactions between the nascent polypeptide chain, membrane lipids, and the membrane protein insertion machinery. Negatively charged residues have a lesser impact on topology, and there are usually no significant differences in their distribution between the cytosolic and extracytosolic loops.

In addition to the positive-inside rule, numerous other factors can influence the orientation of transmembrane helices. These are many and complex, and include but are not limited to: the relative location of the first (or only) transmembrane segment within the greater protein, the glycosylation of extracytoplasmic domains in eukaryotes, and the rate at which any soluble domains fold during translation.

Today, our knowledge of the sequence characteristics of membrane proteins has been exploited in numerous different computer programs that seek to predict membrane protein topology from sequence alone. This is discussed in greater detail in Chapter 19.

4.6.6.2 Internal structural symmetry

Many α -TMPs display an interesting feature in their three-dimensional folds: the presence of internal symmetry in which two portions of the transmembrane helical bundle are structurally homologous and related to each other about a two-fold (180°) rotational axis. If the internal two-fold symmetry occurs along an axis that is perpendicular to the membrane plane, then the structural repeat is *non-inverted* (Figure 4.19). An *inverted repeat* arises when the axis of symmetry lies parallel to the membrane plane, giving the two structurally homologous units opposite orientations within the bilayer (Figure 4.19).

Proteins with more than eight transmembrane helices are the most likely to contain internal symmetry, though the structurally homologous regions are not necessarily related to each other in sequence. Structural symmetry is thought to result from the duplication (and possible inversion at the protein level) of all or a portion of a gene during evolution. This creates two paralogous regions that are fused together to encode a longer, internally symmetric protein.

Internal duplications are frequently observed in secondary active transporters (Section 13.4), where they are thought to assist in the concerted structural switching between different conformational states in the transport cycle. One such example is the protein LeuT, which is discussed in detail in Section 13.4.2.1.

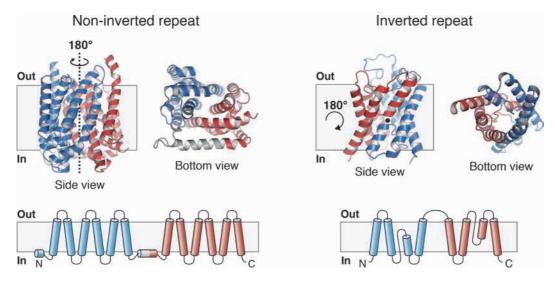


Fig. 4.19 ■ Internal structural symmetry in α-helical membrane proteins. The N- and C-terminal repeating halves of the proteins are colored blue and red, respectively. Left: Two-fold symmetry perpendicular to the membrane plane produces a non-inverted repeat. Coordinates are from the membrane-spanning domain of the bacterial efflux transporter AcrB (PDB: 2GIF). Right: Two-fold rotational symmetry parallel to the membrane plane results in an inverted repeat. The rotational symmetry axis is going into the page (represented by black dot). Coordinates are from the protein aquaporin (2F2B). Each half of the protein also contains a reentrant loop that does not cross the membrane (Section 4.6.4.3).

4.6.7 The Insertion of Helical Membrane Proteins into the Bilayer

4.6.7.1 The two pathways of membrane targeting and insertion

Like soluble proteins, those destined to reside within the membrane begin their lives on the ribosome in the aqueous intracellular milieu (Chapter 11). Ultimately, however, they must be targeted to and integrated into a lipid bilayer. This can occur via one of two independent pathways: the co-translational or the post-translational pathway. The best understood and by far the most common of these is the co-translational pathway, in which membrane incorporation occurs *during* translation. This pathway is used for the insertion of all polytopic membrane proteins and all bitopic proteins except those with a tail anchor (Section 4.4.2 and Figure 4.20). Tail-anchored bitopic proteins — which possess their only membrane-spanning domain within ~30 residues of the C-terminus — are instead incorporated via the post-translational pathway. Membrane insertion via the post-translational pathway takes place after the entire polypeptide has been synthesized (Figure 4.20).

CO-TRANSLATIONAL PATHWAY >30 residues Stop-transfer Signal after first TM domain anchored anchored Polytopic (or signal peptide) Extra-SRP cytosolic Cytosolic RNC complex

POST-TRANSLATIONAL PATHWAY

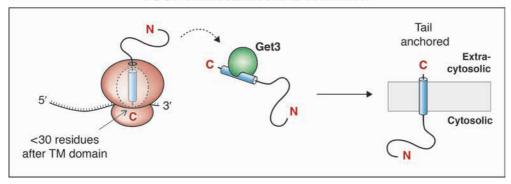


Fig. 4.20 ■ The two pathways of helical membrane protein insertion. The cleavable signal peptide on stop-transfer anchored proteins (Section 4.4.2) is colored yellow. Labeled termini indicate a single possible orientation, whereas unlabeled termini indicate that the protein can exist in either orientation (depending on the internal topogenic signals; Section 4.6.6.1). "TM" stands for "transmembrane". Get3 (TRC40 in humans) chaperones tail-anchored proteins to the membrane for insertion.

To understand why tail-anchored bitopic proteins must be inserted post-translationally, we must first touch upon the early events of membrane targeting via the classical co-translational pathway. This pathway is initiated when either an N-terminal signal peptide (Section 4.4.2) or a transmembrane domain emerges from the ribosomal exit tunnel. Both are hydrophobic in character and are recognized and bound by the signal recognition particle (SRP Figure 4.20), which temporarily arrests translation and chaperones the ribosome-nascent chain complex (RNC) to the membrane for co-translational insertion (Section 4.6.7.2).

The ribosomal exit tunnel itself can accommodate around 30-40 residues before they are exposed to the cytosol and accessible to the SRP. It is therefore evident that for tailanchored bitopic proteins (whose only transmembrane helix lies very close to the C-terminus), translation will have ceased before the membrane-spanning region emerges

from the exit tunnel (Figure 4.20). This simple physical constraint thus necessitates the post-translational insertion of such proteins. And because the SRP binds only very poorly to transmembrane regions once translation is completed, the post-translational pathway also utilizes an entirely different set of proteins for membrane targeting and insertion than the co-translational pathway.

As yet, we are still lacking a detailed understanding of post-translational insertion into the bilayer. We will therefore examine the details of membrane protein insertion in the context of the more common and better-examined co-translational pathway.

4.6.7.2 Details of co-translational membrane insertion

Membrane proteins destined for co-translational insertion are first chaperoned by the SRP to the correct membrane. In prokaryotes this is the inner membrane and in eukaryotes usually the endoplasmic reticulum membrane, from which the protein can be trafficked to other locations after insertion and folding. The SRP then transfers the RNC to the translocon, which is the membrane-spanning molecular assembly that facilitates integration into the bilayer.

The translocon is an impressive multiprotein complex with many different subunits and accessory proteins that are involved in different stages of membrane protein insertion, folding and assembly (and also the export of secreted soluble proteins). At its core is the heterotrimeric Sec complex: Sec $61\alpha\beta\gamma$ in eukaryotes and SecYEG in prokaryotes. The Sec complex forms a pathway across the membrane that houses the nascent polypeptide chain during translation (Figures 4.21 and 4.22). The actual protein-conducting pore is created by the Sec61α (SecY) subunit, which possesses two distinct exit routes for the nascent chain: a vertical channel through which the polypeptide may pass out of the cytosol and into the ER lumen (or periplasm/extracellular space), and a *lateral gate* that opens into the membrane plane (Figure 4.21, top panel). The protein also possesses a plug domain that blocks the vertical channel in the absence of a nascent chain (Figure 4.21).

So how does the translocon actually insert helical membrane proteins into the bilayer, and how is topology established? To answer this question, let us first examine the integration of a signal-anchored bitopic membrane protein. As we saw in Section 4.4.2, signalanchored bitopic proteins can adopt one of either a Type I (extracytosolic N-terminus) or a Type II (cytosolic N-terminus) orientation, and their correct topology must be established at the translocon before insertion. To achieve a Type I orientation, the polypeptide can simply enter the translocon with its N-terminus leading and its C-terminal region (still attached to the ribosome) remaining in the cytosol (Figure 4.21, top panel). Once the transmembrane helix is positioned in the central channel of the Sec complex, the lateral gate opens to allow the nascent chain to partition into the bilayer. This step occurs spontaneously, and is thought to be driven by the preference of a hydrophobic helix for the aliphatic membrane interior rather than the predominantly polar Sec61 channel.

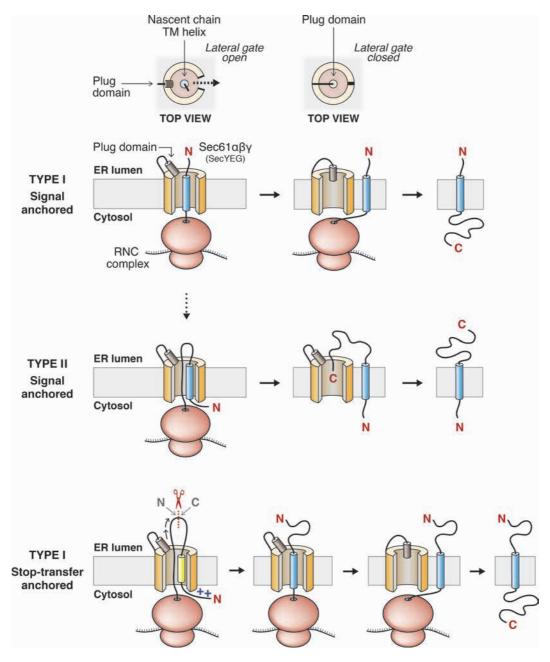


Fig. 4.21 The co-translational insertion of bitopic membrane proteins. A portion of the translocon has been cut away in the front view. "TM" stands for "transmembrane". The eukaryotic ER lumen corresponds to the periplasm or extracellular space in prokaryotes. The dashed arrow between the top two panels represents a possible change in helix orientation while the nascent chain is within the translocon channel. The signal peptide for a stop-transfer anchored protein is colored yellow and the transmembrane domain is blue. Polytopic membrane proteins follow these same steps for the insertion of their first transmembrane helix. See Figure 4.22 for the later stages of polytopic membrane protein insertion.

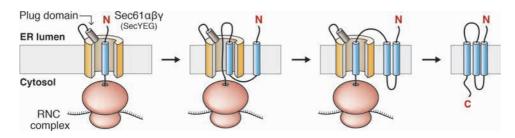


Fig. 4.22 ■ Co-translational insertion of polytopic membrane proteins. In this example, the first transmembrane helix is in a Type I orientation. See Figure 4.21 for the early stages of topogenesis when the first helix is in a Type II orientation.

How a Type II signal-anchored bitopic protein assumes its correct topology within the translocon is perhaps more difficult to imagine. Although the molecular details of this process are still emerging, it appears that the nascent chain might test both Types I and II orientations within the translocon channel (Figure 4.21, top and middle panels). The helix will then partition into the membrane when its preferred orientation is established. Alternatively, the helix might sometimes directly enter the translocon in a Type II orientation (Figure 4.21, middle panel). For example, this could be favored if the presence of a large amino-terminal domain precludes passage of the N-terminus through the channel.

Unlike signal-anchored proteins, stop-transfer anchored proteins — which are targeted to the membrane by a cleavable hydrophobic signal peptide (Section 4.4.2) — are constrained to a Type I orientation. This is because the signal peptide itself possesses a short positively charged region at its N-terminus before the hydrophobic membranetargeting region. Upon its delivery to the translocon, the signal peptide will therefore adopt a Type II orientation consistent with the positive-inside rule (Section 4.6.6.1 and Figure 4.21, bottom panel). It follows that cleavage of the signal peptide will produce a protein with its N-terminus outside of the cytosol and its transmembrane domain (once translated and incorporated) in a Type I orientation (Figure 4.21, bottom panel).

All of the principles described above regarding bitopic proteins can also apply to the topogenesis and insertion of the first transmembrane helix of a polytopic membrane protein. Downstream helices are generally incorporated into the bilayer in the order in which they leave the ribosome, and the topology of each is established sequentially at the translocon (Figure 4.22). It follows that the topology of the first helix will strongly affect those of later helices. This is because each must be incorporated in alternating orientations in order for the polypeptide backbone to move back and forth across the membrane. However, although this sequential model is likely to be true for most polytopic membrane proteins, there are some known examples in which transmembrane helices can remain outside of the bilayer until downstream helices have been synthesized and integrated. Polytopic membrane protein topogenesis is therefore a complex process that must be examined on a case-by-case basis.

β-Barrel Membrane Proteins (OMPs)

4.7.1 Occurrence and Structural Diversity

The second of the two classes of transmembrane proteins are the β -barrels. Unlike members of the α -helical class (Section 4.6), β -barrels reside only in the outer membranes of gram-negative bacteria and in the outer membranes of mitochondria and plastids in eukaryotes. Because of this, transmembrane β -barrels are often called *outer membrane* proteins or OMPs. OMPs are much less common than helical membrane proteins, representing only 2-3% of all genes in gram-negative bacteria and an even lower fraction in eukaryotes. Yet despite their relative scarcity, OMPs perform various critical cellular functions. Often they are porins (Section 13.2.1) or even energy-dependent transporters that facilitate the exchange of molecules across membranes, while others function as enzymes or bacterial adhesins.

The transmembrane regions of all OMPs are formed from a basic β -barrel fold, yet they still display great diversity in their overall structures (Figure 4.23). For example, OMPs may exist as monomers, or they can oligomerize to create functional dimers or trimers. The size of the β -barrel pore can also vary greatly, with those of known structure

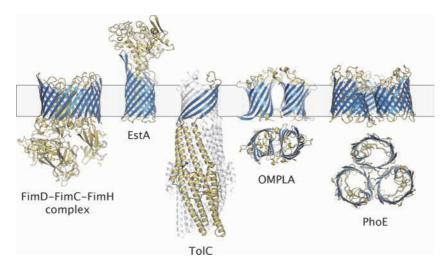


Fig. 4.23 \blacksquare The structural diversity of transmembrane β-barrel proteins. The β-barrel portions of each protein are colored blue and the loop regions, soluble domains, or plug domains are colored yellow. The structures shown are the Fim complex (PDB: 3RFZ; the β -barrel is formed from FimD), EstA (3KVN), TolC (1EK9), OMPLA (1QD6) and PhoE (1PHO). TolC is a homotrimer composed of three identical subunits that form a single β -barrel (one subunit is shown in color and the others in grey). OMPLA and PhoE homooligomerize into dimers and trimers, respectively. The structures of these two proteins are also shown from the top.

possessing between 8 and 24 transmembrane strands in total (though, as we shall see in Section 4.7.3, they are almost always comprised of an even number of strands). Some OMPs are also assembled from multiple chains that each contribute a subset of strands to form a single barrel. Extreme examples of this are the pore-forming toxins secreted by many bacteria, which form giant β-barrels that are assembled from dozens of individual subunits. Finally, OMPs can also contain soluble domains that project away from the membrane, or they can possess plug domains that lie within the barrel interior. Plug domains can have diverse structures and often mediate gating in β-barrel transporters.

4.7.2 Comparison with Soluble β -Barrels

β-barrels are formed from β-sheet secondary structure (Section 2.3.2.1) that is wrapped around in a cylindrical shape until the edge strands hydrogen bond with each other. This fold is not unique to membrane proteins, and many soluble proteins — such as the famous "green fluorescent protein" (GFP) — also form β-barrels. However, there are some important structural differences between soluble and membrane-spanning barrels. Firstly, the β -sheets from OMPs possess an antiparallel β -meander or up-and-down topology. This means that their component β -strands are arranged in the same order in their secondary structure as they appear in the sequence. It follows that the N- and C-terminal strands of the protein will hydrogen bond with each other to close the barrel (Figure 4.24, left panel). Membrane-spanning β -barrels are sometimes referred to as up-and-down barrels, and their topology is distinct from soluble β -barrels that are often assembled from other β -sheet motifs such as the Greek key (Section 3.2.3.1).

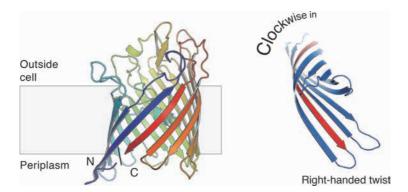


Fig. 4.24 ■ Hallmark structural features of OMPs. *Left*: The antiparallel arrangement of β-strands in an OMP. Coordinates are from the protein OmpF (PDB: 2OMF). Right: The handedness of β -barrels. Because the β -strands move *into* the page when traced in a clockwise direction, the barrel is right-handed (Section 2.3.1.2). Coordinates are from the protein OmpX (1QJ9).

Another contrast between soluble β -barrels and OMPs lies in the different types of amino acids on their inner and outer faces. This arises from the different environments in which the two groups of proteins fold and function — one in the aqueous intracellular milieu, and the other in a hydrophobic membrane. The exteriors of soluble β -barrels consist of polar or charged residues that can interact with the solvent, while their interiors form a densely packed hydrophobic core. This pattern is reversed in integral membrane β -barrels (Section 4.7.4). OMPs are also usually much larger than their soluble counterparts, since they need not maintain a tightly packed hydrophobic interior and instead often form a pathway across the membrane.

For the remainder of this chapter, the broad term " β -barrel" will be used when describing structural features that are shared between the soluble and transmembrane barrels. In contrast, membrane-spanning β -barrels will be referred to strictly as "OMPs".

4.7.3 Basic Structural Features of OMPs

Each membrane-spanning β -strand in an OMP is typically formed from 9–11 residues. But instead of crossing the membrane in a strictly vertical manner, the β -strands of all OMPs (and also soluble β -barrels) are tilted relative to barrel axis (Figure 4.24, left panel). This tilt is caused by the signature "twist" of β -sheets (Section 2.3.2.1), and it also allows side chains to pack favorably within the barrel. β -barrels are always twisted in a right-handed direction. This is evident when the barrel is viewed down its axis (Figure 4.24, right panel). A left-handed twist would result in a barrel that looks like the mirror image of that shown in Figure 4.24 (for a discussion of handedness, see Section 2.3.1.2).

OMPs almost always contain an even number of β -strands. In this way, the N- and C-terminal β -strands will be antiparallel in the assembled barrel and the up-and-down arrangement of the β -strands will be preserved (Figure 4.24, left panel). One notable exception is the eukaryotic voltage-dependent anion channel (VDAC), which is a mitochondrial OMP responsible for the transfer of metabolites between mitochondria and the cytosol. VDAC contains 19 β -strands and is the first known example of an OMP possessing parallel N- and C-terminal strands.

Finally, bacterial OMPs have yet another signature feature: the β -turns between strands on the periplasmic face are tight and the loops short, whereas the extracellular face consists of long loops that are often involved in the protein's function (Figure 4.24, left panel).

4.7.4 Sequence Features of OMPs

Many OMPs exhibit polar interiors that provide a pathway for water-soluble solutes to traverse the otherwise impermeable membrane. This means that OMPs are usually amphipathic proteins — presenting hydrophobic residues (such as valine, leucine and

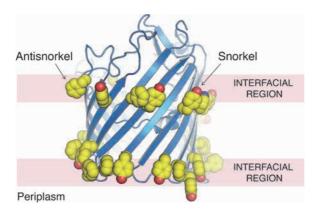


Fig. 4.25 ■ Aromatic girdles in the interfacial regions of OMPs. The side chains of tryptophan, tyrosine and phenylalanine residues are shown as spheres. Note that phenylalanine is not a characteristic of the aromatic girdles of the α -helical membrane proteins (Section 4.6.5.1). An antisnorkeling phenylalanine and a snorkeling tyrosine are labeled. Coordinates are from OmpF (PDB: 2OMF).

isoleucine) to the membrane aliphatic core and possessing polar or charged residues on their inside faces (Figure 4.10). In order to maintain this spatially biased amino acid distribution, OMP strands display a very loose sequence pattern of repeating polar-hydrophobic dyads. Consecutive polar and hydrophobic side chains will therefore lie on opposite faces of the barrel when arranged in a β -sheet (Section 2.3.2.1 and Figure 4.10). The amphipathic nature of OMPs means that it is difficult to identify membrane-spanning strands from sequence analysis alone. For this reason, the computational prediction of OMP topology remains a challenging task that lags well behind progress on their α -helical counterparts (Section 19.2.3.1).

Those portions of OMPs that lie within the interfacial regions of the membrane (Section 4.2) are enriched in the aromatic residues tryptophan, tyrosine and phenylalanine. These residues form two aromatic girdles around the protein (Figure 4.25) a structural feature that was explained in Section 4.6.5.1. Tryptophan, tyrosine, and phenylalanine also display the same snorkeling or antisnorkeling behavior as the α -helical membrane proteins (Section 4.6.5.1, Figure 4.17 and Figure 4.25).

4.7.5 β -Barrel Shear Numbers

The structures of both soluble and membrane-spanning β-barrels can be described almost entirely by two parameters: the *shear number* (*S*, detailed below) and the number of constituent strands (n). Once S and n for a barrel are known, various other structural features may be calculated. These include the overall tilt angle of the β -sheet, the radius of the barrel (assuming circular cross-section), and even the pattern of side chain packing within the barrel.

Because the β -sheet of a β -barrel is tilted with respect to the barrel axis, the component β -strands will also be staggered with respect to each other (Figure 4.26). The degree of stagger is quantified by the shear number, and higher values will signify a greater tilt angle (α) between the strands and the vertical barrel axis (Figure 4.27). The shear number of a barrel depends upon the number of its component strands, and is optimally in the range n < S < 2n to produce ideal tilt angles of around 45°.

But how are shear numbers actually calculated? One simple method is illustrated in Figure 4.27. Start at residue *a* in strand 1 and move from strand-to-strand around the barrel following a path parallel to the hydrogen-bonding direction. Because of the stagger of

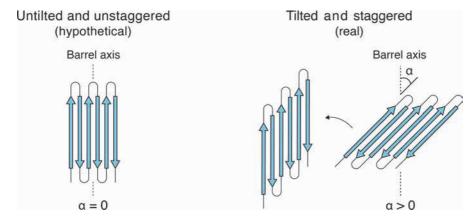


Fig. 4.26 \blacksquare The relationship between sheet tilt and strand stagger. *Left:* A hypothetical barrel with a tilt angle (α) of zero and with unstaggered strands. *Right:* Real β-barrels are composed of tilted and staggered β-strands. The tilted barrel is shown in two orientations to assist comparison with the untilted barrel and illustrate the strand stagger.

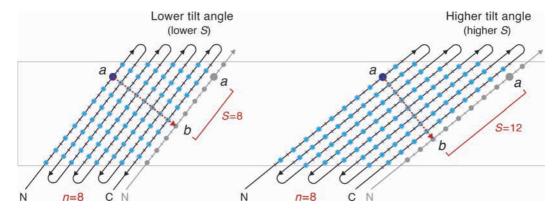


Fig. 4.27 ■ Calculating the shear number of a β -barrel. Two examples are given with different tilt angles. Each sheet has 8 strands in total (n = 8), but the first strand is duplicated in grey. Blue circles represent amino acids facing outwards from the barrel, and dashes those facing inwards.

the strands, one complete revolution will see the path arrive at a different residue in strand 1 (residue b) than the path began. The offset between b and a (that is, b-a), is the shear number. Its value will always be positive because β -barrels are always tilted in the right-hand direction. Shear numbers are also always even numbers, as adjacent amino acid side chains on neighboring β -strands point in the same direction (one of either outward or inward; Figures 2.16 and 4.27).

One complication in calculating the shear number of a real β -barrel can be the occurrence of β -bulges, which introduce extra residues into β -strands (Section 2.3.2.2). The shear number should therefore always be given as the minimum S that is not affected by the presence of β -bulges.

4.7.6 Targeting and Insertion of OMPs

All OMPs are synthesized from beginning to end on cytosolic ribosomes and inserted post-translationally into the membrane. Folding and insertion take place concurrently, and the protein must therefore be trafficked to the correct membrane in an unfolded form.

In gram-negative bacteria, the extended polypeptide must first cross the inner membrane before reaching the outer membrane into which it will be inserted (Figure 4.28, left panel). This is achieved through an N-terminal signal peptide that targets the protein for export to the periplasm. Bacterial OMPs cross the inner membrane via the vertical channel of the SecYEG translocon — the protein complex that also facilitates α -helical membrane protein insertion and that was described in detail in Section 4.6.7.2. In contrast, eukaryotic OMPs possess no signal peptide and are targeted to the outer membranes of mitochondria and chloroplasts via internal signal sequences. Interestingly, mitochondrial OMPs are not inserted directly into the outer membrane from the cytosol. Instead, they

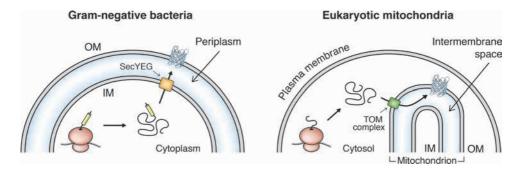


Fig. 4.28 ■ OMP trafficking in gram-negative bacteria and eukaryotic mitochondria. Outer and inner membranes are labeled "OM" and "IM", respectively. The N-terminal signal peptide of bacterial OMPs is represented by a yellow cylinder. Unfolded OMPs pass through the bacterial SecYEG translocon or the eukaryotic TOM complex before folding at the outer membrane. Chaperone proteins have been omitted from the figure for clarity.

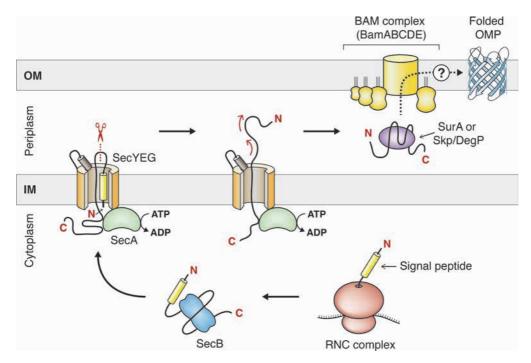


Fig. 4.29 ■ The pathway of bacterial OMP biogenesis and insertion. The SecA protein is responsible for pumping the unfolded polypeptide through the SecYEG translocon. The unfolded protein is bound by different chaperones in the cytoplasm (SecB) and the periplasm (SurA or Skp/DegP) before being delivered to the BAM complex for folding and insertion. The BAM components depicted in the figure are those present in E. coli. The actual mechanism of BAM-mediated folding and insertion is still unknown. Abbreviations are as follows: RNC, ribosome-nascent chain; IM, inner membrane and OM, outer membrane.

first cross the outer membrane in an unfolded form via the TOM complex, and then fold and insert from within the intermembrane space (Figure 4.28, right panel).

The OMPs from gram-negative bacteria are targeted to and cross the inner membrane through the same SecAB pathway used for the post-translational secretion of soluble proteins. A general overview of this pathway is illustrated in Figure 4.29. Once the polypeptide has reached the periplasm, folding and insertion are facilitated by the BAM complex. The precise components of the BAM complex vary between different bacterial species, but in *E. coli* it consists of five subunits: the β-barrel protein BamA and four peripheral membrane proteins attached to the bilayer via lipid-anchors (BamBCDE). BamA is the core of the OMP insertion machinery, and this subunit is the only component of the complex that is conserved across all gram-negative bacteria and also eukaryotic mitochondria. The BamA subunit is absolutely required for in vivo folding and membrane integration of OMPs, which takes place in a concerted manner. However, the precise mechanism by which BamA and its accessory subunits facilitate this process still awaits discovery.

For Further Reading (Section 4.3)

Original Research

Deisenhofer J, Epp O, Miki K, et al. (1985) Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3A resolution. Nature 318: 618–624.

Henderson R, Unwin PN. (1975) Three-dimensional model of purple membrane obtained by electron microscopy. Nature 257: 28–32.

Reviews

Deisenhofer J, Michel H. (1989) Nobel lecture. The photosynthetic reaction centre from the purple bacterium Rhodopseudomonas viridis. EMBO J 8: 2149–2170.

For Further Reading (Section 4.6)

Original Research

Devaraneni PK, Conti B, Matsumura Y, et al. (2011) Stepwise insertion and inversion of a type II signal anchor sequence in the ribosome-Sec61 translocon complex. Cell 146: 134–147.

Granseth E, von Heijne G, Elofsson A. (2005) A study of the membrane-water interface region of membrane proteins. J Mol Biol 346: 377–385.

MacKenzie KR, Prestegard JH, Engelman DM. (1997) A transmembrane helix dimer: Structure and implications. Science 276: 131–133.

Tsirigos KD, Hennerdal A, Kall L, Elofsson A. (2012) A guideline to proteome-wide alpha-helical membrane protein topology predictions. *Proteomics* **12**: 2282–2294.

Van den Berg B, Clemons WM Jr., Collinson I, et al. (2004) X-ray structure of a protein-conducting channel. Nature 427: 36-44.

Walters RF, DeGrado WF. (2006) Helix-packing motifs in membrane proteins. Proc Nat Acad Sci USA 103: 13658–13663.

Reviews

Bowie JU. (2011) Membrane protein folding: How important are hydrogen bonds? Curr Opin Struct

Dowhan W, Bogdanov, M. (2009) Lipid-dependent membrane protein topogenesis. Ann Rev Biochem 78: 515-540.

- Li E, Wimley WC, Hristova K. (2012) Transmembrane helix dimerization: Beyond the search for sequence motifs. *Biochim Biophys Acta* **1818**: 183–193.
- Senes A, Engel DE, DeGrado WF. (2004) Folding of helical membrane proteins: The role of polar, GxxxG-like and proline motifs. *Curr Opin Struct Biol* **14**: 465–479.
- Shao S, Hegde RS. (2011) Membrane protein insertion at the endoplasmic reticulum. *Ann Rev Cell Dev Biol* **27**: 25–56.
- von Heijne G. (2006) Membrane-protein topology. Nat Rev Mol Cell Biol 7: 909-918.

For Further Reading (Section 4.7)

Original Research

- Murzin AG, Lesk AM, Chothia C. (1994) Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. *J Mol Biol* **236**: 1369–1381.
- Kim S, Malinverni JC, Sliz *et al.* (2007) Structure and function of an essential component of the outer membrane protein assembly machine. *Science* **317**: 961–964.

Reviews

- Fairman JW, Noinaj N, Buchanan SK. (2011) The structural biology of beta-barrel membrane proteins: A summary of recent reports. *Curr Opin Struct Biol* **21**: 523–531.
- Knowles TJ, Scott-Tucker A, Overduin M, Henderson IR. (2009) Membrane protein architects: The role of the BAM complex in outer membrane protein assembly. *Nat Rev Microbiol* 7: 206–214.

Basics of Nucleic Acid Structure

Nucleic acids, also known as polynucleotides, are linear polymers composed of nucleotide monomers. The polynucleotide backbone is composed of a backbone of five-carbon sugar units connected through phosphate groups. To each sugar unit a purine or a pyrimidine molecule is attached. These are called *bases*, a term stemming from the acid-base properties of these molecules. There are two main classes of polynucleotides essential to all forms of life, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).

DNA stores the genetic information of the cell in a varying sequence of four bases: adenine (A), thymine (T), guanine (G) and cytosine (C). This sequence is converted into functional protein molecules in the translation process (Chapter 11), using rRNA molecules as temporary copies of specific sections of the DNA. These working copies of genes or groups of genes are called messenger rRNA (mRNA) because they act as messengers between the DNA molecule and the translation machinery, and they also serve as messengers for the metabolic state of the cell (Chapter 10).

rRNA molecules can also store genetic information, for example, in some viruses, but rRNA has important roles in other processes as well. The most central metabolic processes involving nucleic acids are replication, transcription and regulation. They all depend upon the recognition of the nucleic acid sequence information by proteins. Thus, it is important to understand the structural properties of DNA and rRNA in order to see the possibilities for biologically relevant interactions between the nucleic acids and proteins.

On April 25th, 1953, three papers appeared back-to-back in the Journal, *Nature*. Two of them contained the experimental data and results of diffraction experiments on DNA fibers. The first of the three — and the most famous — is the short paper by Watson and Crick with the title "A Structure for Deoxyribose Nucleic Acid". At the time it was known that DNA consists of nucleotide units connected through their phosphate groups, but the overall conformation of the so-called sugar-phosphate backbone was unknown. Watson and Crick did not carry out any experiments of their own, but they did get information from X-ray fiber diffraction photographs resulting from experiments carried out by the authors of the other two papers, most notably the famous "Photo 51" (Figure 5.1) recorded by R.G. Gosling, a collaborator of Rosalind Franklin. Photo 51 was a fiber diffraction

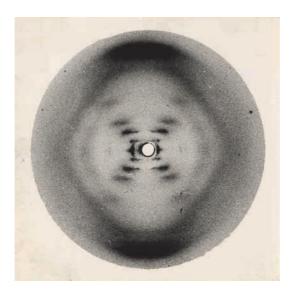


Fig. 5.1 ■ The famous "photo 51" fiber diffraction image of B-DNA taken by R.G. Gosling. The image was created by exposing the stretched, wet DNA fibers to X-rays at an angle perpendicular to the fiber axis. The "X" pattern is clearly visible. The spots in the "X" are seen to fall on horizontal, equidistant lines also known as *layer lines*. The distance between these layer lines can be computed to be 34 Å, corresponding to the periodicity of the helix fiber. At the bottom and the top of the photo, a very intense reflection can be seen. These reflections fall on the tenth layer line corresponding to an interbase distance of 3.4 Å. From these facts it can be deduced that the DNA fiber is helical with a repeat distance of 34 Å, with 10 bases per full turn of the helix. (Reproduced with permission from Franklin RE, Gosling RG. (1953) Molecular configuration in sodium thymonucleate. *Nature* 171: 740–741. Copyright (2009) Macmillan Publishers Ltd.)

image from "wet" DNA fibers, called "B-DNA", and was simpler than the semicrystalline diffraction patterns of "A-DNA" which were more complicated and difficult to interpret.

The theory of diffraction from a 3-dimensional helical array of atoms had been worked out by Crick and Cochran in 1952, and their theory predicted that the X-ray diffraction would exhibit a characteristic "X" pattern, and that it could be described mathematically by a combination of Bessel functions. Having established the mathematical basis of helical diffraction, Crick was able to recognize that the X-ray diffraction images recorded by Rosalind Franklin and Gosling revealed a helical structure of the DNA. Furthermore, the width of the helix (20 Å), the distance between base pairs (3.4 Å) and the pitch of helix (height of one complete turn of 34 Å) could be calculated from the diffraction pattern, revealing the repeating structure along the axis of the fiber. Perhaps the most ground-breaking discovery of Watson and Crick was the formation of base pairs in the interior of the helix. In their own words:

"The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fiber axis.

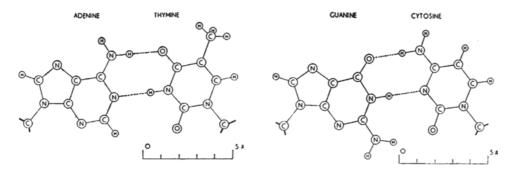


Fig. 5.2 ■ The base pairing of adenine with thymine and guanine with cytosine from Watson and Crick's paper in *Nature*, May 30, 1953. It was later realized that there are three hydrogen bonds in the G:C pair. (Reproduced with permission from Watson, JD and Crick, FHC. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967. Copyright (2009) Macmillan Publishers Ltd.)

They are joined together in pairs, with a single base from the other chain, so that the two lie side by side with identical *z*-coordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur."

This base-pairing scheme (Figure 5.2), which was worked out by Watson using card-board models of the bases, was a key step in the development of the DNA model, because it presented important restraints on the structure: the phosphate backbones are placed on the outside of the DNA fiber, and the number of chains is two. From the monoclinic symmetry of the diffraction images of A-DNA it was concluded that the two chains had to be antiparallel. Thus, armed with the essential facts originating from experiments carried out by the groups of Rosalind Franklin and Maurice Wilkins, plus knowledge of the stereochemical properties of nucleotides from previous crystallographic work, Watson and Crick were able to build a physical model of DNA using brass atomic model components made in the workshop of the MRC laboratory at Cambridge.

The model showed two sugar-phosphate chains running in opposite directions (Figure 5.3). The bases are on the inside of the helix and the phosphates on the outside. Given the geometric restrictions of the nucleotides in their standard conformation, the planes of the bases are perpendicular to the helix axis, and two grooves appear spiraling along the fiber, one wider than the other. Because the A:T and G:C base pairs have the same dimensions, the helix is a cylinder with a uniform diameter, regardless of the sequence of the bases. There are 10 base pairs per complete turn of the helix, thus there is an angle of 36° between adjacent residues in the same chain.

Looking at the paper with today's eyes, it is unlikely that at paper of 128 lines, with no experimental data, and with only 6 references, would be accepted for publication in *Nature*. However, Watson and Crick's model for DNA explained so many scientific results that were not understood at the time that their model simply could not be dismissed. The experiments carried out by Avery, MacLeod, and McCarty a decade earlier had suggested that genetic

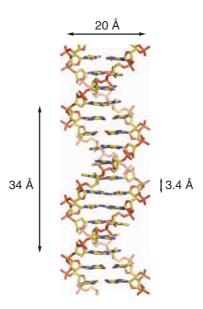


Fig. 5.3 ■ The DNA double helix. The sugar-phosphate backbone forms two ridges or grooves winding around each other. The base pairs are at the center "arranged like a pile of pennies" as Wilkins *et al.* write in their paper in *Nature*, April 25th, 1953.

information is indeed carried by DNA, but noone had any notion how. The most widespread idea was that genes were copied by some kind of template mechanism, and the first models of DNA, for example, that proposed by Pauling, had the bases sticking outwards with the sugar-phosphate backbone sitting in the interior of the fiber. The idea was that the pattern of bases could somehow be recognized and copied by other components of the cell, perhaps involving proteins. In their paper, Watson and Crick inconspicuously write:

"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

Watson and Crick elaborated on this laconic sentence in a second paper in the May 30th edition of *Nature*. They established that the first feature of the double helix DNA structure is that it consists of two chains, held together by hydrogen bonds between the bases, and that one member of the pair must be a purine and the other a pyrimidine in order for the two chains to be held together. If a pair consisted of two purines there would be insufficient room for it. The bases in the two strands are in their most probable tautomeric form, which means that the only base paring possible is adenine with thymine and guanine with cytosine, see Figure 5.2. The consequence of this arrangement is that DNA becomes a pair of templates, and each chain can act as a template for the formation of a new companion chain. Watson and Crick proposed that free nucleotides will join up by forming hydrogen bonds to one of the bases on the chain already formed. They also realized that it is essential for the chains to untwist if they are to separate, and that for the

entire chromosome to separate, a considerable amount of uncoiling would be necessary. They were also able to explain the phenomenon of spontaneous mutation. The base pairing scheme explained the experiments of Chargaff, who had established earlier that the ratios of adenine to thymine and guanine to cytosine were both 1:1.

All of the observations in Watson and Crick's second paper are essentially correct, but it would be many years before the double helix structure was confirmed by X-ray crystallography. This happened in 1974, when the structure of transfer rRNA (tRNA) was determined independently by two groups: one at MIT, Boston the other at the MRC Laboratory of Molecular Biology, Cambridge. tRNA indeed contains double helical stems, as will be seen later in this chapter. The first detailed view of a right-handed B-DNA double helix came in 1980, from structural studies of the so-called Dickerson-Drew dodecamer, a self-complimentary oligonucleotide CGCGAATTCGCG. A year earlier, however, the X-ray crystal structure of the self-complimentary CGCGCG hexamer was determined by the A. Rich group at M.I.T., but the structure displayed a left-handed helix and was termed Z-DNA. Between the discoveries of the left-handed Z-DNA structure and the right-handed Dickerson-Drew B-DNA dodecamer there was an anxious time where scientists began to wonder whether the Watson-Crick double helix was indeed correct.

The discovery of the DNA double helix more than anything marks the birth of the field of molecular biology. Before this, many life processes of the cell were known, but very poorly understood. Most studies in life science were carried out on bacterial cells and bacteriophages, and how these observations related to pure biochemistry was not at all well understood. The view held today, that life processes of the cell are due to a large number of macromolecules, essentially nanomachines, that interact with each other and smaller molecules along strict chemical principles, was far into the future in 1953. The model of the DNA double helix was the first glimpse of that future.

5.1 **Building Blocks**

The building blocks of DNA and rRNA are the nucleotides, a cyclic ribose (RNA) or 2'-deoxyribose (DNA) pentose sugar with a phosphate group attached to the 5'-hydroxyl, and a base attached to the 1' carbon atom of the ring.

In the polynucleotide, the phosphate of one nucleotide is connected to the 3'-hydroxyl group of the sugar of the next nucleotide (Figure 5.4). The nucleic acid molecules thus have a 5' and a 3' end. The sugar with the attached base is called a nucleoside, and in the cell we find a variety of nucleosides. The conformation of the backbone is defined by a number of angles, called α , β , γ , δ , ε and ζ (Figure 5.4).

Ribose moieties are found in rRNA and 2'-deoxyribose moieties in DNA (Figure 5.5). The carbon atoms in the pentose sugar moiety are numbered from 1' to 5'. The only

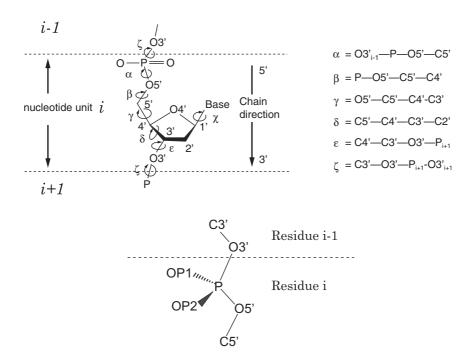


Fig. 5.4 ■ Section of the DNA backbone, showing the atom naming and the naming of various torsion angles. The cyclic ring is the deoxyribose moiety of the backbone, and successive ribose units are connected via phosphate groups. The backbone is therefore often referred to as "the sugar-phosphate backbone". *Bottom*: The direction of the chain is by convention defined by two oxygen atoms on the ribose, O5′ and O3′, and specifies the direction in which transcription takes place (from 5′ to 3′ in the new strand).

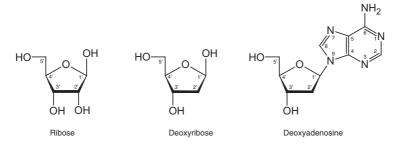


Fig. 5.5 ■ *Left and middle*: The two types of riboses found in nucleic acid molecules. *Right*: A nucleoside, deoxyadenosine, showing how the adenine base is attached to the sugar, in this case, 2′-deoxyribose.

difference between ribose and 2'-deoxyribose is that the 2'-hydroxyl group in the latter is missing. The 2'-hydroxyl group contributes to the catalytic potential of rRNA molecules. From the polymer perspective, the 3'- and 5'- carbon positions are the most important because they are directly involved in the phosphodiester linkages, linking the monomers together.

The ribose and deoxyribose molecules are not flat, but have a pucker. Several types of puckers are possible, but the most common ones are the C2′-endo conformation, found in double-stranded DNA molecules, and the C3′-endo conformation, which is the most common conformation in double-stranded RNA. Single-stranded regions in rRNA molecules can have both types of puckers. Puckers are divided into two main categories, E and T. If four atoms of the aldofuranose ring lie approximately in a plane, the conformation is described as envelope (E), due to the pictorial resemblance to an envelope, otherwise the conformation is described as twist (T) (Figure 5.6). The conformation of the ribose is the main difference between the A- and B-forms of DNA. In the A-form, the ribose is in the C3′-endo conformation, and in the B-form it is in the C2′-endo conformation.

The puckering of the ribose is conveniently described by the phase angle P, which is defined in terms of the internal ribose conformation angles $\nu_0 - \nu_4$ (see Figure 5.7)¹:

$$P = \arctan \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_0)}{2\nu_2 \left[\sin(\pi/5) + \sin(2\pi/5)\right]}$$

$$^{2}E \qquad C5' \qquad B \qquad C5' \qquad B \qquad ^{3}E$$

$$^{2}T_3 \qquad C5' \qquad C5' \qquad B \qquad C5' \qquad B \qquad ^{3}T_2$$

$$E_3 \qquad C5' \qquad C5' \qquad B \qquad C5' \qquad B \qquad E_2$$

$$C5' \qquad C5' \qquad C5' \qquad B \qquad C5' \qquad B \qquad E_2$$

$$C5' \qquad C5' \qquad C5'$$

Fig. 5.6 ■ Various sugar ring puckering conformations. Those on the left are denoted S (for south); those on the right, N (for north). The C3′-endo conformation is seen at the top right, and the C2′-endo conformation at the top left. The notation of E- and T- conformations is also given. Superscript numbers preceding E or T refer to carbon atoms on the same side of the reference plane (horizontal line) as C5′. Subscripts following E or T denote atoms on the opposite side of the reference plane.

¹The definition of the pseudorotation is given by Altona C and Sundaralingam MJ. (1972) *Am Chem Soc* **94**, 8205–8212. When using the definition of torsion angles given in Figure 5.7, the conversion given on page 8207 of the paper must be used, so the formula given here looks different from their Equation (3).

$$v^4$$
 v^0 v^0 $v^0 = C4' - 04' - C1' - C2'$
 $v^1 = 04' - C1' - C2' - C3'$
 $v^2 = C1' - C2' - C3' - C4'$
 $v^2 = C2' - C3' - C4' - 04'$
 $v^2 = C3' - C4' - 04' - C1'$

Fig. 5.7 ■ Naming scheme for the torsion angles of the sugar ring in nucleotides.

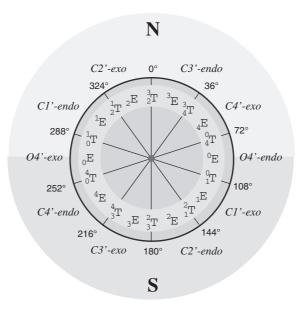


Fig. 5.8 • Diagram showing the correlation between the phase angle P and the ribose conformation. Twist conformations (T) arise at values of P which are even multiples of 18°, and envelope (E) conformations arise in between. Conformational angles of P are divided into two categories, north (N) which are conformations with a *positive* value of ν_2 , and south (S) which are conformations with a *negative* value of ν_2 .

P must be in the interval 0–360°, so if ν_0 < 0° then P = P + 180°. The various conformational nomenclatures of the ribose ring can conveniently be drawn along a circle as a function of P (Figure 5.8).

Adenine and guanine are the two most common purine bases, but inosine is found in some nucleic acid molecules. DNA has two types of pyrimidines, cytosine and thymine. In RNA, thymine is normally replaced by a similar base, uracil, which lacks the methyl group found in thymine (Figure 5.9).

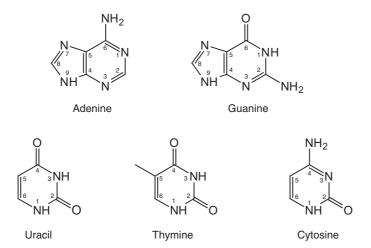


Fig. 5.9 ■ The most common bases found in nucleic acids: the top row are purines; the bottom row pyrimidines. The atom-numbering scheme of purines and pyrimidines is given.

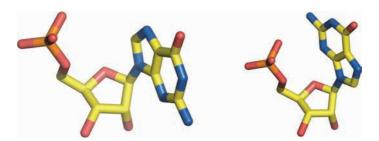


Fig. 5.10 ■ The *anti* (*left*) and *syn* (*right*) conformations of guanine monophosphate.

There are two preferred ways of arranging the base in relation to the ribose, *syn* and *anti* (Figures 5.10 and 5.11). In pyrimidine nucleotides, only the *anti* conformation is found, because this avoids collision between the oxygen and the ribose. Purines can have both orientations, but the *anti* conformation is the most common of the two.

Mammalian DNA is known to contain a methylated base 5-methylcytosine (m⁵C). Bacterial DNA contains this one and two other methylated bases, namely, N6-methyladenine (m⁶A) and N4-methylcytosine (m⁴C). In bacteria, the main function of m⁵C and m⁴C is protection against its own restriction nucleases, so that these destroy only foreign non-methylated DNA and not the organism's own hereditary material. The methylated adenine m⁶A is thought to be involved in the regulation of virulence and the control of several bacterial functions, such as the replication, repair, expression and transposition of DNA.

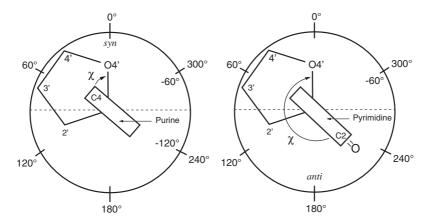


Fig. 5.11 • Diagram defining the torsion angle χ around the N-glycosidic bond. The pentagon illustrates the ribose unit, and the base is seen edge-on. The sequence of atoms chosen to define this angle is O4′–C1′–N9–C4 for purine and O4′–C1′–N1–C2 for pyrimidine derivatives. Thus when $\chi = 0^{\circ}$ the O4′–C1′ bond is eclipsed with the N9–C4 bond for purine and the N1–C2 bond for pyrimidine derivatives. The *syn* conformation is defined as $\chi = \pm 90^{\circ}$ and *anti* as $\chi = 180 \pm 90^{\circ}$; thus, the *syn* conformational region is given by the upper half-circle, and the *anti* conformation by the lower half. The purine on the left is therefore in the *syn* conformation, and the pyrimidine on the right, in the *anti* conformation.

5.2 DNA Secondary Structure: Double Helix

DNA can exist as a single chain or a double helix. The basis of life is the ability of nucleic acid molecules to form base pairs, where the sequence of bases of one nucleic acid chain is bound to another chain through base pairing. A single chain can be used as a template for the synthesis of another with the complementary sequence. The direction of a polynucleotide can be defined by looking at the phosphodiester linkages between adjacent nucleotides (see Figure 5.4). In the sugar-phosphate part, the phosphate groups connect to the 3′- carbon of one deoxyribose moiety and the 5′ carbon of the next moiety, thereby linking successive deoxyriboses together. The two ends of a chain differ; the end where the 5′- carbon is not connected to another nucleotide is called the 5′ end. The other end is called the 3′- end. The two ends may or may not have free phosphate groups.

Each of the four bases in DNA has a unique set of hydrogen bond donors and acceptors that allows it to form base pairs with the other bases. In double-stranded DNA we have AT (adenine-thymine) base pairs with two hydrogen bonds and GC (guanine-cytidine) base pairs with three hydrogen bonds (Figure 5.12). These interactions are called Watson–Crick base pairs to honor the scientists who first suggested that these base pairs are the basis of heredity.

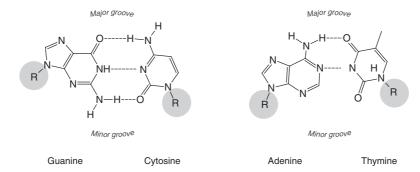


Fig. 5.12 ■ The Watson–Crick base pairs. The sugar moieties are represented by R. Notice that the GC base pair on the left interacts via three hydrogen bonds, whereas the AT base pair on the right has only two. This makes the GC base pair and thus GC-rich DNA more stable than the AT base pair and AT-rich DNA.

In a double helix, the flat bases that base pair in the center are perpendicular to the helical axes. The two chains, which form a cylindrical spiral, run in opposite directions. The 5' end of one chain is paired with the 3'- end of the other strand, and vice versa. This can be represented with an arrow for each chain, running from the 5'- end to the 3'- end. The arrows point in different directions — they are antiparallel. The orientation parameters are crucial during the copying of a single chain/strand

In an ideal DNA double helix, there are approximately 10 base pairs per turn of DNA. Usually, the helical conformation is right-handed, that is, it twists to the right as do the threads on most screws. The sugar-phosphate backbone of the two antiparallel chains forms ridges on the edges of the helix. There are two grooves between the ridges formed by the ribose-phosphate backbone. The two grooves are of different widths and thus traditionally called the major and minor grooves. The major groove is wider and the bases are more accessible than in the minor groove. The exposed edges of the base pairs contain different hydrogen bond donors and acceptors, which can interact with protein molecules. For each base pair, four different combinations of donors and acceptors exist at the exposed edges in the major and minor grooves. This creates a variety of patterns along the DNA double helix, available for highly specific DNA-protein interactions (see Chapters 10 and 11).

Two factors are mainly responsible for the stability of the double helix structure: the base pairing between complementary strands and — especially — the stacking between adjacent base pairs. DNA molecules show considerable variation in conformation, dependent on the actual base sequence, the AT/GC content, and the presence of counter ions and stabilizing/destabilizing proteins. Since the AT base pair only contains two hydrogen bonds, it is easy to distort. AT-rich sequences are functionally important, and often serve as binding sites of DNA-binding proteins. In the crystal structure of the TATA box binding protein complex, the protein induces a significant bending of the DNA (see Section 10.5.1.2).

Double-stranded DNA can adopt several conformations, also called *forms* (Figure 5.13). These conformations are determined by the activity of water and the nature of counterions. *In vivo*, the conformation is also influenced by the presence of proteins. Even if various forms have been observed under *in vitro* conditions, it is still not known whether they play any physiological role under *in vivo* conditions.

In aqueous solutions, the DNA helix most often occurs in the B-form, which has 10 bases per turn. One full turn measures 34 Å in the axial direction. The helical axis runs through the center of each base pair and the bases are stacked perpendicularly to the axis.

In vitro, under moderate relative humidity, a structural transition occurs and DNA adopts the A-form. This form is more condensed and stabilized by helix-to-helix interactions. One could imagine that this form is also found *in vivo* under the conditions where naked DNA is tightly packed, for example, in some viruses. In A-DNA, the base pairs are not perpendicular to the axis but tilted by 13 to 19°. The helical axis is shifted from the center of the base pairs and it runs through the enlarged major groove, while the minor groove has shrunk. Hence, the exposure of the bases is different than in the B-form and the sequence specific interactions between the protein and DNA change in character.

When the water activity, under *in vitro* conditions, becomes even lower, the C-form may occur. This form is less condensed and less crystalline and it may occur during predenaturing conditions. Several other forms of DNA have also been reported. One of the most drastic conformations has been described for shorter DNA with alternating G and C bases. This form, Z-DNA, is left-handed, with a repeat of a dinucleotide unit. It has six dinucleotides per turn and exhibits a characteristic zigzag backbone (Figure 5.13, right). Table 5.1 presents a summary of the most important helix parameters.

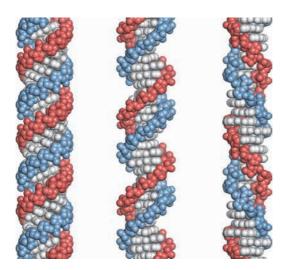


Fig. 5.13 ■ CPK representations of A-, B- and Z-DNA.

| | A | В | Z |
|------------------------------------|----------|-----------------|---------------------------|
| Base pairs per turn | 11 | 10 | 12 |
| Rotation per base pair | 33° | 36° | -30° |
| Distance between base pairs | 2.55 Å | $3.4~{\rm \AA}$ | 3.7 Å |
| Tilt of base pairs | 19° | -1° | -9° |
| Diameter of helix | 23 Å | 20 Å | 18 Å |
| Sugar pucker | C3'-endo | C2'-endo | C2'-endo (C), C2'-exo (G) |
| Configuration of N-glycosidic bond | anti | anti | anti (C), syn (G) |

TABLE 5.1 Helical Parameters of the Three Main Conformations of DNA

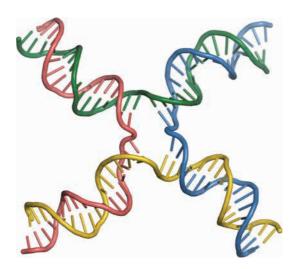


Fig. 5.14 ■ A Holliday junction (PDB: 3CRX).

Several structures show mixtures of various different geometries, likely reflecting the real situation in the cell. Recently obtained data on biologically relevant structures, such as an X-shaped Holliday junction structure, have been very informative. This conformation is observed in DNA undergoing recombination in which the two double helices are held together at the crossing over site. During this process, the two helical regions exchange strands. This is the molecular background for much of the biological variation (Figure 5.14).

A fundamental question in structural biology is how sequence defines conformation. This is a very complex task for proteins and RNA, which can adopt a variety of conformations. DNA is more rigid and can, in general, adopt a more limited number of structures. It has long been a dream that if one determines all possible structures of shorter sequences, these modules together would explain the sequence effects on the overall structure. This would help to predict the structure of longer molecules. However, reality is often more complex than what can be predicted from small modules.

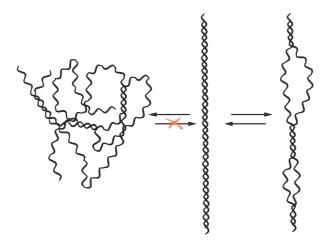


Fig. 5.15 ■ Schematic illustration showing the transition between a double helix conformation, a local, reversible denaturation and extensive, irreversible denaturation. Under conditions of complete denaturation, random base pairing may occur because there is always a large number of short complementary regions.

DNA adopts a helical conformation in room temperature aqueous solution. If the temperature is raised, the helix undergoes a helix-coil transition and the molecule becomes a random coil (Figure 5.15). This process is also called denaturation or melting, and can be reversible. A helix-coil transition may be cooperative or non-cooperative. In the fully cooperative case there is an all-or-nothing character. In other words, any given molecule is either in the completely helical state or in the random-coil state. Studies of the denaturation properties are essential to understand the biochemical aspects of DNA. In short, reversible, local denaturation is an essential element of replication and gene expression.

5.2.1 Deviation from Ideality

A DNA double helix is not a perfectly regular structure and does not form a perfect rod. In contrast, DNA in chromosomes adopts a tightly wound and very densely packed structure. The parameters defining the conformation of the bases in a double helix have been defined in the Cambridge convention, drawn up by R.E. Dickerson. These parameters do not take the conformation of the backbone into consideration, only the relative spatial orientation of the purines and pyrimidine moieties. The six inter-base parameters (shift, slide, rise, tilt, roll, twist) describe the local conformation of two neighboring bases along the backbone. Of these, the first three are purely translational, and the latter three are rotational. In practice, the inter-base conformation is a linear combination of these six parameters.

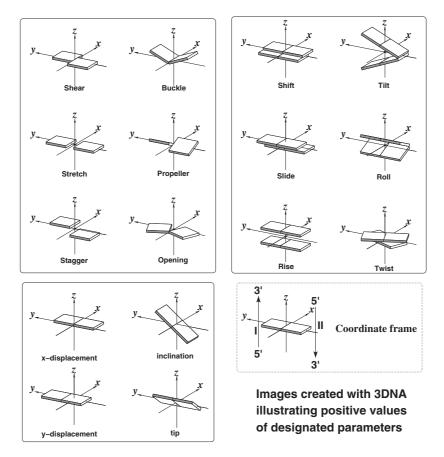


Fig. 5.16 ■ The Cambridge convention's definition of the standard parameters used to describe the relative orientation of bases in a double helix. (Reprinted with permission from Lu X-J and Olson WK. (2003) 3DNA: A software package for the analysis, rebuilding, and visualization of three-dimensional nuclei acid structures. Nucl Acids Res 31: 5108-5121. Copyright (2003) Oxford University Press.)

Base pairs are formed between bases on opposite strands and, in B-DNA base pairs, are not perfectly planar. The deviation from perfect planarity depends on the participants in the base pair. For example, A:T base pairs typically have a propeller twist of approximately -15 to -20°. The Cambridge convention defines six parameters that describe the deviation from planarity of base pairs within a double helix (shear, stretch, stagger, buckle, propeller, opening). These parameters describe the translational and rotational displacement between participant bases in a base pair. The parameters are calculated in a standard coordinate frame, defined by the helix axis (see Figure 5.16).

When making a double helix from base pairs of different values of the inter-base pair parameters, the path of the helix depends on those values. Figure 5.17 shows different double helix conformations due to variation of two of the parameters. The double helix will adopt a conformation that maximizes the van der Waals interaction between bases.

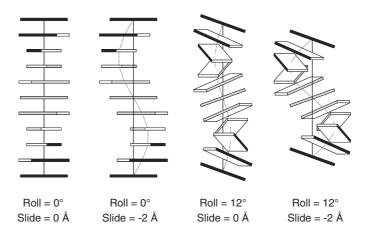


Fig. 5.17 ■ Effect of the inter-base parameters on the helical path. (Reprinted with permission from Lu X-J and Olson WK. (2003) 3DNA: A software package for the analysis, rebuilding, and visualization of three-dimensional nuclei acid structures. *Nucl Acids Res* **31:** 5108–5121. Copyright (2003) Oxford University Press.)

Some combinations of parameters are forbidden due to steric clashes, whereas others can give rise to helix bending and turning. Whether a conformation is allowed or not is also influenced by the exact sequence of the DNA. The genetic code is degenerate, so it is possible to vary the base composition of the DNA without causing a mutation in the protein. Thus, the DNA-base sequence not only bears genetic information, but also information pertaining to its own three-dimensional structure.

5.3 Tertiary Structure of RNA

5.3.1 Structural and Functional Differences Between RNA and DNA

Although rRNA is chemically very similar to DNA, their three-dimensional structures are very different. Most DNA molecules are large and in the form of double helices, rRNA molecules are composed of shorter helical regions with base pairs and single-stranded regions where the bases and the backbone interact in different ways. And whereas the DNA almost exclusively displays the classical Watson–Crick double helix structure, rRNA exhibits a wide range of structural folds. Some rRNA molecules have a defined, unique conformation that is crucial for their function, while other molecules may have a more

flexible conformation. The tRNA molecules (see Section 5.3.10.1 and Chapter 11) and the rRNA molecules in ribosomes (Figure 5.18) are examples of rRNA with an ordered conformation. The coding region of messenger rRNA is an example of a flexible rRNA that must continuously pass through the ribosome, and local secondary structures need to be unzipped (or bypassed) during the coding process. The molecules have double helical regions that are joined using various types of interactions where single-stranded regions interact with sequentially remote double helical segments.

Chemically, the difference between DNA and rRNA appears negligible. Whereas DNA is constructed from deoxyribonucleotides, rRNA consists of ribonucleotides, the difference being a single oxygen atom attached to the 2' carbon of the ribose sugar ring. Another difference is in the selection of bases: rRNA makes use of uracil instead of thymine, which are identical except for the methyl group in the thymine. But this seems odd; why not use the same four bases in both DNA and RNA? The answer comes from evolution as well as chemistry. The chemical degradation of cytosine to form uracil is one of the most common

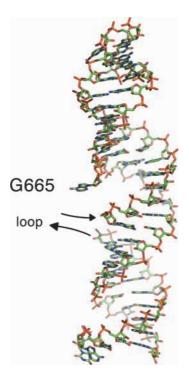


Fig. 5.18 ■ RNA (helix 23 in 16S rRNA from *T. thermophilus* ribosomes, PDB: 1N32). Nucleotides 656 to 750 are shown except nucleotides 717 to 733 (marked "loop") and 683–704 (connecting the strands at the bottom of the drawing). Nucleotide 656 is at the top of the drawing and the chains go all the way to nucleotide 682 at the bottom, with the base of G665 not stacking (marked). Nucleotides 705 to 750 go from the bottom to the top. Double-stranded RNA is normally close to the DNA A-conformation (see Section 5.2).

mutations of DNA, but it can be easily recognized and repaired. If DNA made use of uracil instead of thymine, the cell would not know which uracil bases to repair. It is important for the cell to protect the integrity of the DNA because that is what is passed onto future generations.

Another important difference is that rRNA does not have a complementary strand and most often is single-stranded. However, rRNA often contains stretches of selfcomplementary sequences that enable it to fold back on itself, forming stretches of "hairpin" loops that can adopt a double helical structure. rRNA structure is often thought of as consisting of helices (stems) and loops, and a hairpin structure is often referred to as a "stem-loop" structure. The helical regions of rRNA often have the A-conformation, with 11 base pairs per turn. The ribose has C3'-endo pucker and the major and minor grooves have a shape that differs from that of DNA. Watson–Crick base pairs are found in the double-stranded regions of rRNA molecules, as are many other types of base pairs or arrangements of the bases. The reason for this is that the sequence of rRNA molecules does not contain perfectly matching sequences. The tendency to form double helices with base stacking leads to base pairs with fewer hydrogen bonds and less optimal backbone conformation, for example, GU pairs with two hydrogen bonds but with the bases shifted compared to the normal Watson–Crick base pairs.

Biologically, rRNA performs several distinct roles. It acts as a carrier of genetic information (mRNA, viral genomes in rRNA viruses), and as a structural entity (rRNA). It has a role in recognition (tRNA, siRNA), and in catalysis of chemical reactions (ribozymes). These diverse roles stem from the ability of rRNA molecules to adopt a wide range of three-dimensional structures, which again rise from the fact that rRNA is more flexible than DNA. But how can rRNA be more flexible than DNA? One would think that the ribose group, with an extra hydroxyl group, would be sterically more restrained. The answer is that while the deoxyribose prefers a 2'-endo sugar pucker conformation, the ribose prefers a 3'-endo conformation. This results in a conformation of the sugar-phosphate backbone that is more linear, and thus more able to adopt different conformations. So, while the ribose sugar ring is undoubtedly rigid, it still allows for a greater flexibility of the rRNA backbone.

This difference in flexibility is also manifested in the double helical conformation of RNA. The altered sugar puckering shortens the distance between adjacent phosphates by about 1 Å (Figure 5.19). While DNA prefers the B-helix form with 10 nucleotides per turn, rRNA prefers the A-form with 11–12 nucleotides per turn. In DNA, the base pairs are centered over the helical axis. In an rRNA double helix, the base pairs slide ~5 Å away from the helical center. All these factors contribute to the tighter packing of the rRNA double helix.

The surface of an rRNA helix is also quite different from the DNA double helix. The major groove of rRNA is very narrow and deep, accentuated by the fact that rRNA does not have the thymine methyl group, which resides in the major groove. In contrast, the minor groove is wide and shallow. For this reason, the major and minor grooves in rRNA are more descriptively referred to as the deep groove and the shallow groove.

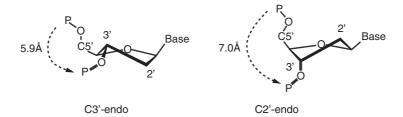


Fig. 5.19 ■ The two types of sugar pucker most commonly found in nucleic acids. The C3′-endo pucker is prevalent in RNA and A-form DNA, whereas the C2′-endo pucker is characteristic of B-form DNA. It is seen that the C3′-endo pucker produces a significantly shorter phosphate-phosphate distance in the backbone, resulting in a more compact helical conformation.

The deep groove of rRNA is a favored binding site of cations, water molecules and protein side chains.

5.3.2 Primary, Secondary and Tertiary Structure

We can classify the different structural levels of rRNA in the same way as with proteins. The secondary structure is determined by base pairing. In addition to the normal Watson–Crick base pairs, mismatches such as GU base pairs are found. The secondary structure can form complicated patterns like the cruciform of tRNA. The secondary structure is formed by hydrogen bonds, with an energy of ~-12kJ/mol. The formation of just a single base pair tends to be energetically unfavorable. The stacking of base pairs gives the largest stabilizing energy contribution of ~-23kJ/mol. The stacking energies of base pairs are not symmetric, i.e. stacking of GC on AU is not the same as the stacking energy of AU on GC. The most complex structural level is the tertiary one. The tertiary structure is formed by the spatial organization of the secondary structural elements, i.e. the stem-loop regions. A loop may form hydrogen bonds to another part of the structure, thus stabilizing it. In addition, van der Waals and electrostatic forces contribute to tertiary structure formation, and finally protein molecules can interact with, and stabilize rRNA tertiary structure. A fundamental principle is that the rRNA chain, like most polypeptides, is unable to form a topological knot.

Large rRNA structures are typically determined to resolutions of at best 2.5 Å. This is likely caused by the experimental difficulties in obtaining material of sufficiently high purity for production of high quality crystals. At medium resolution (2.5–3.5 Å), the phosphates and base planes can be located quite reliably, but ribose rings, and in particular their exact conformations, are typically not well defined in the electron density. Until recently, only limited structural information was available on rRNA, due to the experimental difficulties encountered in rRNA crystallography. Recent years, starting with the landmark achievements of the ribosome structures, have brought a flurry of crystal structures of rRNA molecules at increasingly higher resolutions.

5.3.3 rRNA Allows Alternative Base Pairing

Many of the bases in structural rRNA molecules have been observed to participate in non-Watson–Crick base-pairing interactions. These bases are often involved in forming and stabilizing the three-dimensional structure. A certain class of these alternative base pairings is named after Karst Hoogsteen, who first suggested their existence in 1963.

Although alternate base pairings sometimes have fewer hydrogen bonds than the canonical Watson–Crick base pairing, enhanced base stacking can compensate for the lower stability and become energetically favorable. rRNA is thus able to adapt to various structural requirements while maintaining optimal stability. Sites with alternative base pairing are often sites of biological importance, because the unsatisfied hydrogen bonds are able to take part in interactions with a protein or other factor.

In an attempt to rationalize non-canonical base pairing, Leontis and Westhof proposed a scheme where the base pairing edges on purines and pyrimidines are defined. Purines have three such edges [Watson–Crick (WC), Hoogsteen (H) and sugar (S) edges] and pyrimidines have two (Figure 5.20). A base pair is categorized according to the edges participating in the base pairing. A standard base pair is thus called WC/WC, and a Hoogsteen base pair can be, for example, H/WC or WC/H. With three edges on purines, and two on pyrimidines, there are six combinations, and if parallel/antiparallel chain direction is taken into account, there are a total of 12 different base pairing families into which all base pairs can be categorized (Table 5.2). A base pair is called *cis* if the glycosidic bonds are on the same side of a line of symmetry through the plane of the base pair, and *trans* if they are on opposite sides.

In the classical definition, Hoogsteen base pairing generally always involves an interaction along the "Hoogsteen edge" of purines, which has two hydrogen-bond acceptors on guanine (N7 and O6), and an acceptor and a donor on adenine (N7 and N6). In reverse Hoogsteen base pairing, the two chains are parallel, meaning that the pyrimidine partner is flipped, and the ribose rings end up in the *trans* position.

Figure 5.21 shows a standard Watson–Crick GC base pair (*cis* WC/WC) compared to the GC reverse base pair (*trans* WC/WC) that can be obtained between parallel backbone

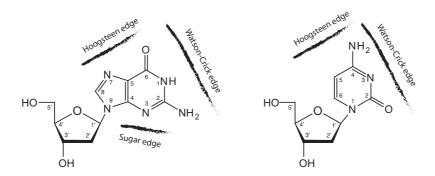


Fig. 5.20 ■ Definition on the base pairing edges on purines and pyrimidines.

| No. | Glycosidic Bond Orientation | Interacting Edges | Local Strand Orientation | Symbol |
|-----|--------------------------------|---------------------------|-----------------------------|--|
| 1 | Cis | Watson-Crick/Watson-Crick | Antiparallel | • |
| 2 | Trans | Watson-Crick/Watson-Crick | Parallel | -0- |
| 3 | Cis | Watson-Crick/Hoogsteen | Parallel | •= |
| 4 | Trans | Watson-Crick/Hoogsteen | Antiparallel | ОП |
| 5 | Cis | Watson-Crick/Sugar Edge | Antiparallel | •+ |
| 6 | Trans | Watson-Crick/Sugar Edge | Parallel | O₽> |
| 7 | Cis | Hoogsteen/Hoogsteen | Antiparallel | - |
| 8 | Trans | Hoogsteen/Hoogsteen | Parallel | -0- |
| 9 | Cis | Hoogsteen/Sugar Edge | Parallel | ■→ |
| 10 | Trans | Hoogsteen/Sugar Edge | Antiparallel | $\Box\!$ |
| 11 | Cis | Sugar Edge/Sugar Edge | Antiparallel | + |
| 12 | Trans | Sugar Edge/Sugar Edge | Parallel | → |

TABLE 5.2 Leontis/Westhof-Base Pair Classification of 12 Main Families of Base Pairs

Notice that the symbols are filled in the cis configuration and hollow in the trans configuration. Circles represent Watson-Crick base pairing, squares represent Hoogsteen base pairing, and triangles represent the sugar edge.

Fig. 5.21 ■ *Left*: Canonical Watson–Crick GC base pair (*cis*). *Right*: GC reverse Watson–Crick base pair (trans).

strands. Since parallel strands do not result from the usual stem-loop folding pattern of rRNA, reverse Watson-Crick interactions are tertiary interactions.

Another fact, apparent in three-dimensional structures of rRNA, is that the distance between backbones is different in the cis and trans configurations (Figure 5.22). The distance between backbone strands is shorter in the AU Hoogsteen pair than in the AU reverse Hoogsteen pair.

The term "wobble" base pairing was proposed by Francis Crick to account for the noncomplementary GU base pairings observed in codon-anticodon interactions, manifested

Fig. 5.22 ■ Left: AU Hoogsteen base pair. Center: AU reverse Hoogsteen base pair. Right: AU reverse Watson–Crick base pair. The blue dashed line shows the line of symmetry used to define the cis/trans conformation of the base pair. The AU Hoogsteen base pair is thus cis H/WC and the AU reverse Hoogsteen is trans H/WC.

Fig. 5.23 ■ *Left*: GU wobble. *Right*: GU reverse wobble.

in the degeneracy of the genetic code (see Chapter 11). Figure 5.23 shows the GU wobble base pair, which is one of the most common alternative base-pairing patterns, and the GU reverse wobble, where the uracil group is simply flipped around the axis of the amine hydrogen bond. The GU wobble base pairing results in the loss of a hydrogen bond from the guanine, but the vacant amino group often forms hydrogen bonds to other bases nearby, perhaps in concert with the neighboring imino group. The GU wobble base pairings can be viewed as a canonical Watson–Crick pattern, with a shift of the pyrimidine partner.

The GU wobble pair has geometric properties that enables it to fit very well into a regular A-form helix, and therefore frequently substitute for regular Watson–Crick pairs. In certain chemical environments, the A base can become protonated at the N1 position. When this happens, the A base is able to form hydrogen bonds with a C base, forming an (A+):C base pair. The geometry of the (A+):C base pair is *isosteric* with the GU wobble pair.

5.3.4 Base Triplets and Quadruplets: Prominent Tertiary **Structural Motifs**

The alternative base pairing patterns described in the previous section lead to a rich selection of multiple-base interactions. Whenever a base pair is formed from two nucleotides, several hydrogen donors or acceptors are still available for alternate interactions to occur, either from the amino acid side chains of proteins, or from other nucleotides. The most important of these multi-base interactions are base triples, important in maintaining the tertiary structure of rRNA molecules. Several examples of triple-base interactions will be presented in the following sections.

The possibility of triple-base interactions were first realized in 1957 by Felsenfeld, who demonstrated that a (poly-A):(poly-U) duplex molecule could interact with a second poly-U strand to create a triple strand complex (Figure 5.24). The additional poly-U strand interacts with the major groove of the duplex by forming Hoogsteen base pairs with the poly-A strand of the duplex. Later, it was found that several other sequence combinations can lead to triple helix formation, as long as they contain two pyrimidines and one purine, for example, C:G-C. triple helix formation has been proposed to have a role in gene repair.

Another higher-order structure is the quadruplex or tetraplex. Four guanine residues can associate to form G-tetrads, planar cyclically hydrogen bonded structures (Figure 5.25). If a nucleic acid sequence has a run of two or more guanine bases, four

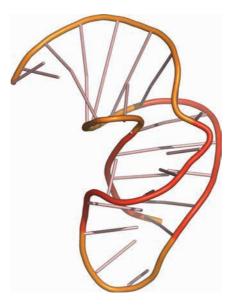


Fig. 5.24 ■ Triple RNA helix from a frame-shifting pseudoknot (PDB: 1e95).

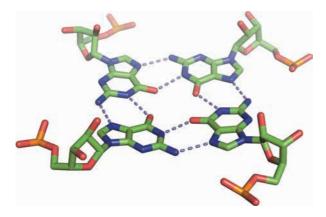


Fig. 5.25 ■ Four guanine residues forming a G-tetrad by four GG Hoogsteen base pair interactions.

separate strands can combine, forming a parallel tetrameric quadruplex (sometimes called tetraplex). Another type of quadruplex is formed when two G-hairpins combine. Finally, if a sequence contains four separated guanine tracts, an intramolecular antiparallel quadruplex can form. These different quadruplex forms are shown in schematic form in Figure 5.26, and one of them in more detail in Figure 5.27. Quadruplexes exhibit an unusual dependence on specific metal ions for their formation and stabilization. A wide variety of cations is able to induce quadruplex structures, including NH₄, Tl⁺, Sr²⁺, Ba²⁺ and Pb²⁺.

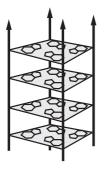
DNA sequences capable of forming quadruplexes are abundant in the telomeres (protein-DNA complexes that cap the ends of the linear chromosomes; Section 9.3).

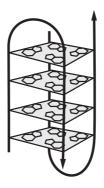
There are several examples of telomeric proteins that bind to quadruplex structures. Another protein shown to interact with DNA quadruplexes is human DNA topoisomerase I. This enzyme can bind to both four-stranded quadruplexes and unimolecular quadruplexes, and can in addition induce the formation of four-stranded quadruplex structures.

Quadruplex-forming sequences have also been found in immunoglobulin switch regions and gene promoter regions. Furthermore, a quadruplex-forming site has been identified in the mRNA of insulin-like growth factor II, where it is thought to preserve a flanking cleavage site or perhaps to repress and to regulate translation of genes in the vicinity.

5.3.5 rRNA Contains Modified Bases

Post-transcriptional modification of rRNA results in an even greater diversity of modified bases than in DNA. This is especially true in functional rRNAs such as tRNA and rRNA, where both ribose and base can be modified. The modified bases can profoundly change the chemical characteristics of the rRNA molecule and can contribute to the stability of the molecule as well as partake in its external interactions and reactions.





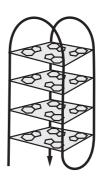


Fig. 5.26 ■ Three basic forms of quadruplex structures. *Left*: Parallel tetrameric quadruplex. Middle: A two G-hairpin quadruplex, one of several types. Right: Intramolecular antiparallel quadruplex.

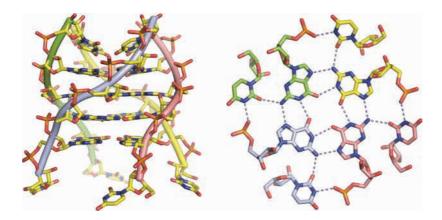


Fig. 5.27 ■ *Left*: An RNA quadruplex — a parallel quadruplex formed by the sequence UGGGGU. Right: An octad of guanine and uracil bases from the same structure. The uracil bases come from neighboring quadruplexes in the crystal (PDB: 1J8G).

A telling example of the importance of post-translational modification is given by an E. coli isoleucine-specific tRNA lle, specific for the codon AUA. This tRNA contains a modified base, lysidine, at the first position of the anticodon. The base lysidine is a cytidine that has been post-translationally modified by the addition of the amino acid lysine to the C2 position (Figure 5.28). If the lysidine residue is replaced by the native cytidine, a marked reduction of isoleucine incorporation is the result, and surprisingly, the appearance of methionine-accepting activity. How can this be? It turns out that in E. coli, the cognate isoleucyl tRNA synthetase recognizes and charges tRNA lle only if the lysidine residue is present in the anticodon loop. Furthermore, the anticodon sequence of tRNA^{Ile} is CAT, which normally codes for methionine. If lysidine is absent, the tRNA will instead be recognized and mischarged by the CAT-recognizing methionyl tRNA ligase.

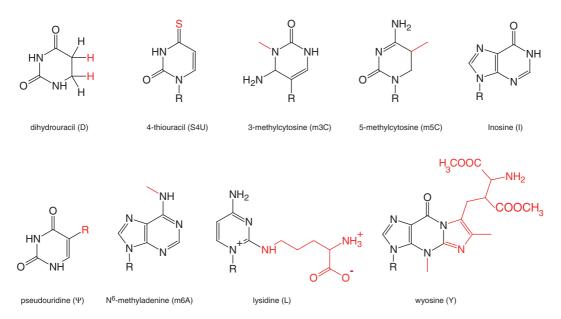


Fig. 5.28 ■ Examples of modified bases in RNA. Modifications are marked in red. R stands for ribose. Notice that not all modified bases are planar.

So a single post-translational modification is responsible for both the codon and amino acid specificity of this tRNA.

5.3.6 Compensating-Base Mutations Reveal rRNA Secondary Structure

The primary structure of the same rRNA molecules from different species can vary extensively just as in the case of proteins. Their structures, however, vary much less. The stretches of sequences that can pair to form helices is a major element in rRNA structure (Figure 5.29). Frequently, there are alternative solutions to the base-pairing problem and this leads to the risk of forming different and conflicting helical arrangements.

The sequence conservation of double helices is frequently less than for single-stranded regions, since the single-stranded regions are often used for specific functions. This has been illustrated in tRNA and ribosomal rRNA. This is a major problem in the prediction of rRNA double helices. However, the conserved structures require that the formation of base pairs in the double helices be conserved. A change on one strand leads to the corresponding change on the other strand. Thus, the prediction of the secondary structure requires information from multiple related sequences.

The secondary structure of rRNA can be accurately determined from alignment of multiple related rRNA sequences from a large number of organisms. The underlying

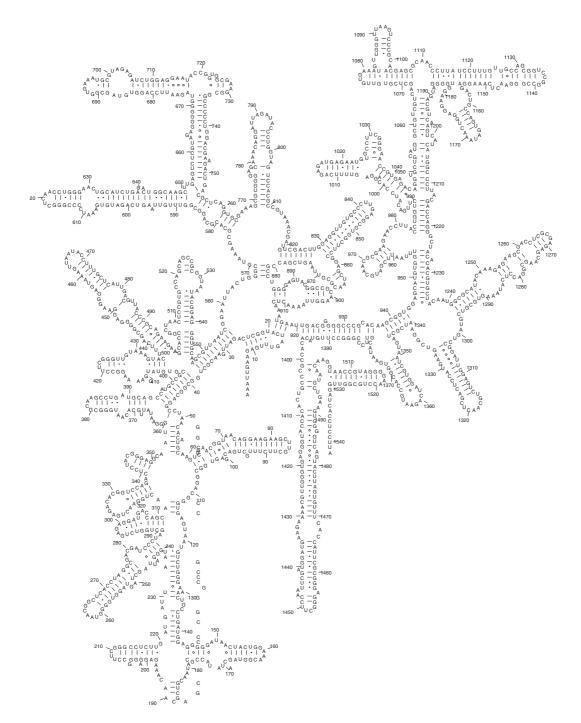


Fig. 5.29 ■ The secondary structures of large RNAs can become very complex. This is the 1542 nucleotide long 16S rRNA from *E. coli*.

assumption is that although mutations in the rRNA can occur, the secondary and tertiary structures of the rRNA are maintained. If a point mutation occurs in a double helical region, the helix will lose a base pair and thus be destabilized. Such mutations will normally be compensated by another base change on the opposite strand, so that the helical base pairing is maintained. Therefore, by looking for compensating base changes in the alignment, it is possible to deduce areas that are able to form double helices invariantly.

Once the secondary structure of the rRNA has been deduced, the same principle can be employed to detect tertiary structural interactions when these are involved in a Watson–Crick base-pairing interaction. However, when tertiary interactions in rRNA are non-canonical (as they very often are) or if the interaction involves a stretch of sequence that for other reasons is conserved, it is not possible for such covariation to be observed. For this reason, determination of tertiary interactions in rRNA is still a theoretical challenge.

5.3.7 rRNA Structural Motifs

The secondary structural motifs of rRNA provide its structural framework in the same way as protein molecules are composed of a variety of combinations of alpha helices and beta sheets. There are two major classes of these motifs, hairpin loops and internal loops. Hairpin loops arise when a stretch of the rRNA sequence is inversely self-complementary. For example, the sequence from the A loop of 23S rRNA (see Chapter 11) is:

GGCUGGCUGUUCGCCAGCC

The first seven base pairs are inversely complementary to the last seven, so the sequence is capable of forming a hairpin loop. Such hairpin loops, or stem-loop structures, are classified according to the number of bases in the unpaired loop region, and are called triloops, tetraloops, pentaloops, etc. (Figure 5.30). One of the factors contributing strongly to the formation of hairpin loops is the stacking force. In the example above, one might think that the structure would assume a pentaloop structure, but in fact the stacking forces motivate the double helix to continue with the formation of the alternative CU base pairing. NMR studies have confirmed that the ribosomal A loop is indeed a triloop.

Unpaired bases tend to propagate the stacking of the double helix, so after the number of unpaired bases in the loop region, hairpin loops are classified according to how those unpaired bases are arranged relative to the double helix, i.e. whether the bases are *looped-out* or *stacked-in*. There are, for example, several classes of triloops, with three, two, one or no bases in the main stack. Furthermore, it is often specified if the stacking-in occurs in the 5' or 3' stack. In special cases where two hairpin loops contain complementary-base sequences, they can form a tertiary interaction, romantically named "kissing loops" (Figure 5.31). This structural motif has been identified in 23S rRNA and as a structural element in the HIV-1 genome.

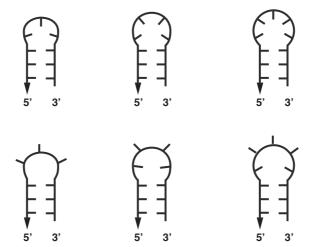


Fig. 5.30 ■ Various hairpin loops. *Top row, from left*: Triloop with two bases in the main stack, tetraloop with four bases in the main stack and pentaloop with four bases in the main stack. *Bottom row, from left*: The same with looped-out bases — triloop, tetraloop with two bases in the main stack, pentaloop with two bases in the main stack (one in the 5′ and one in the 3′ stack).



Fig. 5.31 ■ Schematic representation of "kissing loops".

Another example of kissing loops exists in the crystal structure of tRNA^{Asp} that has the anticodon GUC. Here, two tRNA molecules interact via their anticodon loops in a duplex involving a UU mismatch in the middle position. This duplex has been demonstrated to exist in solution. Other cases of tRNA species with kissing hairpin duplexes that can form in solution are yeast tRNA^{Asp} (GUC) and *E. coli* tRNA^{Val} (GAC) complex. In these, the tRNAs have complementary anticodons.

The rRNA backbone has six degrees of freedom for each residue whereas the polypeptide backbone only has two. This extreme flexibility allows single-stranded rRNA regions to adopt a wide range of conformations. Nevertheless, a couple of single-stranded rRNA motifs are particularly common. The first is the *S-turn*, where an S shape is formed by two consecutive bends in the phosphate-sugar backbone and distinguished by inverted sugar puckers. The S-turn motif is found in the ribosomal loop E motif and the sarcin-ricin loop (see Section 11.4). The other important single-stranded rRNA motif is the *U-turn*, which is a sharp bend in the backbone between the first and second nucleotides, followed by a distinctive stacking of the second and third nucleotides. Hydrogen bonds between the first and third residues often stabilize the motif. U-turns are typical of GNRA loops and are also found in the TΨC loop of tRNA.

5.3.7.1 The GNRA tetraloop

Tetraloops are a particularly common motif in rRNA structures. An especially well-known case of this hairpin loop is the GNRA² loop motif, which closes the hairpin of many rRNA stem-loop regions. The SCOR database lists more than 600 examples of the GNRA motif. The GNRA loop commonly adopts a specific three-dimensional structure, called the *GNRA fold*, with one base in the 5′ stack and three in the 3′ stack. The fold contains a U-turn, stabilized by a sheared GA non-canonical base pair. The GNRA loop motif is often closed by a CG Watson–Crick base pair.

The GNRA fold exists with 4, 5 or 6 nucleotides in the loop. An example of a pentaloop exhibiting the GNRA fold (consensus sequence GNRNA) is GAAAA. In this case, the fourth base is looped out, while the remaining bases adopt equivalent spatial positions as in the GNRA tetraloop (Figure 5.32). The tetraloop family of sequence UMAC forms the same fold (M = [AC]). Similarly, the UNCG motif is a stable tetraloop found in ribosomal and other functional RNAs. In the *UNCG fold*, the U and C bases are part of the 5′ stack, the G is the part of the 3′ stack, and the N base is looped out. The same fold has been observed in a GUUA tetraloop.

Another well-known tetraloop motif is ANYA which is, for example, found in coat-protein binding hairpin loops from MS2 viral rRNA (see Chapter 18).

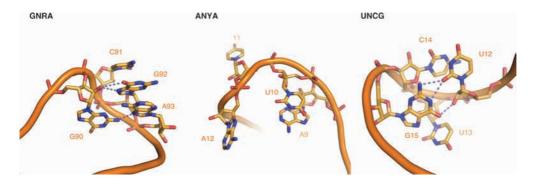


Fig. 5.32 ■ Three-dimensional structures of various tetraloop folds. *Left*: GNRA loop from 5S rRNA (PDB: 1JJ2). The first G in the loop stabilizes the loop by hydrogen bonding to the fourth member. *Middle*: ANYA loop from MS2-RNA complex (PDB: 1DZS). Bases one and two form a stacking interaction, while bases three and four of the loop are looped out and poised to interact with other species. Base 11 is a rare modified base: pyridin-4-one. *Right*: UNCG tetraloop from 16S rRNA (PDB: 1BYJ). The first U and the last G in the tetraloop interact via hydrogen bonds, while bases one and two in the loop form a stacking interaction. The third base in the loop is available for interaction with other species.

²R stands for puRine; N stands for aNy; Y stands for pYrimidine.

5.3.7.2 Internal loops

In general, internal loops adjoin two regular A-form helices, and where the bases of the loop are unpaired, they can often participate in non-canonical base pairing (Figure 5.33). Tightly bound water molecules stabilize the loop, and the water-mediated hydrogen bonds widen the deep minor groove.

Internal loops are often the binding site of proteins. For example, the HIV-1 Rev protein binds at a 10-nucleotide asymmetric internal loop.

Internal loops are classified according to the base-pairing characteristics of the adjoining helices, and the stacking of the bases in the loop itself. Evidently, this leads to a great number of different loop conformations (Table 5.3).

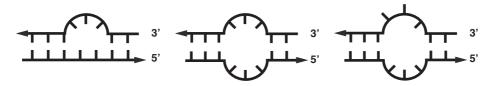


Fig. 5.33 ■ Various examples of internal RNA loops. *Left*: A three-base bulge. *Middle*: A symmetric internal loop with looped-in bases. Right: An internal asymmetric loop with looped-out bases.

| TABLE 5.3 | Internal Loop | Categories | According to | the SCOR | Classification |
|-----------|---------------|------------|--------------|----------|----------------|
| IADLL 3.3 | mitchiai Loop | Categories | According to | me beon | Classification |

| Internal Loops | Subclasses |
|--|------------|
| Stacked, fully paired non-Watson-Crick double strand | 9 |
| Stacked, one base unpaired, flanked by base | 2 |
| Looped-out bases | 2 |
| Loops with base triples | 3 |
| Loops with dinucleotide platforms | 3 |
| Loops with trans-orientation of the glycosidic bonds | 4 |
| Loops with unpaired, unstacked, looped-in bases | 24 |
| Loops with cross-strand stack | 6 |
| Loops with stacked, interdigitated bases | n/a |
| Loops with interrupted stack | n/a |
| Loops with external stacked bases | 2 |
| Kink turn | n/a |
| Helical bending | 4 |
| Loops with two independent stacks | n/a |
| S turn | 2 |
| Loops with a potential base pair not formed | n/a |

The existence of subclasses means that the motif can be categorized in greater detail.

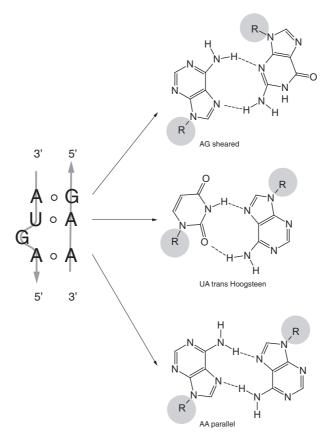


Fig. 5.34 ■ Consensus loop E family motif. Circles in the secondary structure diagram on the left indicate non-Watson-Crick base pairs.

An example of an internal loop that has gained a lot of attention is the loop E motif, first recognized in the 1980s when it was found that similar, conserved internal loops in eukaryotic 5S rRNA and in PSTV viral rRNA shared a surprising disposition for cross-linking when exposed to UV light (Figure 5.34). Later, the loop E motif was found in the sarcin-ricin loop (SRL) of ribosomal 23S rRNA, which is involved in the binding of elongation factors EF-Tu and EF-G. In E. coli 5S rRNA, loop E is known to constitute a specific binding site for ribosomal protein L25 (Figure 5.35). It was therefore apparent that this motif was an important site of activity and molecular recognition. Both 5S rRNA and the SRL have been intensively studied by both high resolution NMR and crystallography (Section 11.5.2.1).

Loop E is an asymmetric internal loop characterized by a highly conserved stack of seven non-Watson-Crick base pairs. The hydrogen-bonding pattern is particularly well conserved. The first base pair is an AG sheared pair, followed by a UA trans Hoogsteen pair, a bulged G base, and finally a trans Hoogsteen AA base pair with locally parallel backbone conformation (Figure 5.34). The AU reverse Hoogsteen base pair observed in loop E is the most abundant AU interaction in the ribosomal rRNAs after the normal Watson-Crick base pair.

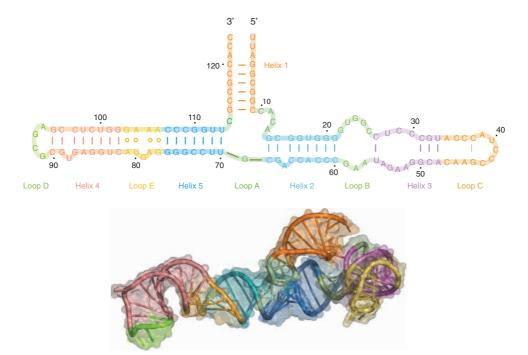


Fig. 5.35 ■ *Top*: Secondary structure of *H. marismortui* 5S rRNA observed in the crystal structure of large ribosomal subunit. *Bottom*: Schematic representation of the three-dimensional structure of 5S rRNA. The orientation and color coding is similar to the figure above.

5.3.7.3 Bulges

A number of insertions in an otherwise regular helical region form a structural motif known as a *bulge*, but in reality a bulge is a special case of an internal loop (Figure 5.36). A bulge inserts a bend in an rRNA double helix and is an important structural element because of its ability to orient stem loops and thus contribute to the overall tertiary folding of the molecule.

The unpaired nucleotides in rRNA bulges can be looped-out or stacked-in. When looped-out, the nucleotides often form sites of tertiary interactions, or specific recognition sites for proteins. Figure 5.36 shows a one-base bulge, but there can be several nucleotides in a bulge.

One particularly well-studied rRNA bulge exists in the trans-activation response region (TAR) of HIV-1 (Figure 5.37). This is a 59-nucleotide rRNA stem-loop structure in the 5′-non-coding region of all HIV-1 mRNAs, which plays an important role in regulating HIV gene expression. The TAR region interacts with the viral *Tat protein* that activates viral gene expression and is obligatory for virus replication. There is experimental evidence indicating that Tat binds directly to the TAR element, and while this binding is

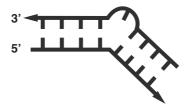


Fig. 5.36 ■ An RNA looped-in bulge motif.

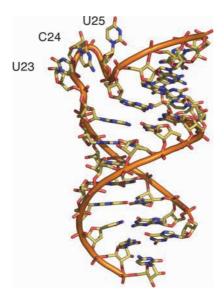


Fig. 5.37 ■ Close-up of the bulge from the TAR stem-loop structure (PDB: 397D).

independent of the nucleotide sequence it depends on the integrity of the upper stem, containing a 3-nucleotide bulge.

5.3.7.4 Junctions

Junctions are regions of rRNA that connect two or more stems. The strands between the stems can have a length ≥ 0 of bases and are called linking or joining regions (Figure 5.38). Three-stem junctions are abundant in stable rRNA, for example, ribosomal rRNA, viral rRNA and ribozymes. Later in this chapter we shall discuss the hammerhead ribozyme, which is a three-stem junction. The four-stem junction — sometimes called a cruciform junction — is also common and is found in tRNAs, for example.

The topology of three-stem junctions has been particularly well described. Three-stem junctions with two helices coaxially stacked can be categorized in three families, depending

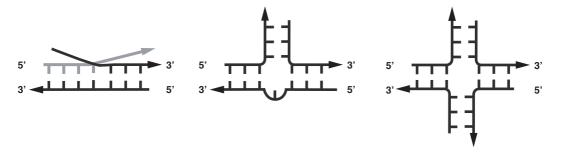


Fig. 5.38 ■ Schematic of two-, three- and four-stem RNA junctions.

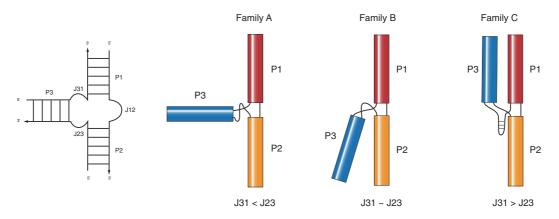


Fig. 5.39 ■ Left: The nomenclature used for three-stem junctions. Right: Schematic drawing of the three families of three-way junctions.

on the length of the linker regions, which is again manifested in the relative orientation of the three stems. The length of the linker regions is significant in situations where the stems pack against each other, because the helices can rotate relative to each other if the linker is long. In Figure 5.39, a schematic representation of the three families of three-stem junctions is shown. In family A, the linker region J31 is shorter than J23 and helix P3 is approximately perpendicular to the P1-P2 coaxial stack. In family B, the linkers J31 and J23 are of equal length and helices P2 and P3 pack alongside each other. Family A and B junctions are found in ribosomal 23S and 16S rRNAs. Finally, in family C, linker J31 is longer than J23, helix P3 packs parallel to helix P1 and linker J31, which often is a hairpin structure, makes extensive interactions with the shallow groove of P2. Family C junctions are found in various structured rRNAs such as hammerhead ribozyme, the P4-P6 domain of self-splicing intron, and signal recognition particle domain S- and G-riboswitches. The spliceosome is described in some detail in Chapter 10.

5.3.8 Xrn1 Resistant rRNA

A fascinating illustration of how a three-stem junction exists and has a three-dimensional structure with functional implications is seen in the rRNA of mosquito-borne flaviviruses (Chapter 18). Many of these viruses are serious human pathogens and emerging world-wide health threats. The virus' genetic material contains *subgenomic flaviviral RNAs* (sfRNAs), discrete rRNA sequences associated with the cytopathicity of the virus. During infection, sfRNAs are produced by partially resisting degradation of the virus' genomic rRNA by the host cell Xrn1, a 5'-3' exonuclease otherwise capable of degrading a wide range of structured rRNAs.

Biochemical and bioinformatics studies on isolated Xrn1-resistant stem-loop structures (xrRNA) showed them to be compactly folded RNAs with highly conserved nucleotides that are crucial for both structure and function, and prompted further structural analysis. Figure 5.40 shows the structure of an xrRNA from the 3′ untranslated region of the Murray Valley Encephalitis virus, determined by X-ray crystallography. The core three-stem junction of the xrRNA structure is not independently folded, but appears when tertiary base pairs are formed. The result is a ring-like structure in which the 5′ end of the rRNA passes through.

This suggests a mechanism where the hosts cell's Xrn1, when working along the genomic rRNA of the flavivirus from the 5' end, encounters a "knot" on the rRNA. Modeling studies suggest that the ring-like structure of xrRNA blocks the enzyme at the entrance to the active site, preventing further unwinding of the viral rRNA by Xrn1's helicase-like activity.

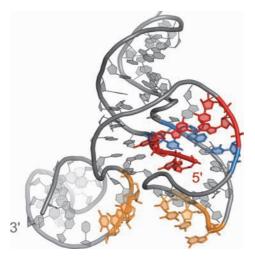


Fig. 5.40 ■ Schematic structure of an xrRNA. The 5′ end is shown in red, and can be seen emerging from the back side of the structure through the center of ring-like structure, which is held together by the base pairing of highly conserved nucleotide sequences shown in in red, orange and blue. In this view, the Xrn1 exonuclease approaches from the front side and encounters a "brace" on the RNA sequence, preventing its progress. (PDB: 4PQV).

5.3.8.1 K-turn

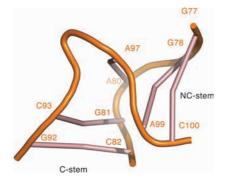
An analysis of the three-dimensional structure of the 50S ribosomal particle from *Haloarcula marismortui* led to the identification of a small rRNA motif, called the *kinkturn* or *K-turn*. The K-turn is a two-stranded, loop-helix motif comprising 15 residues (Figure 5.41). It is an asymmetric internal loop characterized by a kink in the sugarphosphate backbone that causes a sharp turn in the rRNA helix. The bend occurs on the shallow groove side and brings together the shallow grooves of two adjacent helices. One of these, the *C-stem* (canonical stem), comprises only Watson–Crick base pairs, while the other helix is composed of non-canonical base pairs and is therefore called the *NC-stem*.

The two stems are held together by a series of stacking interactions including an A-minor interaction. One of the unpaired loop nucleotides protrudes significantly, making this motif ideal for recognition by ribosomal proteins. There are six kink turns in *H. marismortui* 50S, two in *T. thermophilus* 30S, and one each in U4 snRNA and L30e premRNA. A consensus sequence motif predicts the presence of K-turns in many rRNAs, including the 5′-UTR of L10 mRNA, helix 78 of *E. coli* 23S rRNA and human RNase MRP. The upper part of the consensus K-turn has two CG base pairs and the lower has two AG base pairs normally followed by two Watson–Crick base pairs. The kink generally has three bases on one strand that are not base paired.

One of the unpaired loop nucleotides of the K-turn protrudes significantly. This makes the K-turn an important RNA-recognition motif for proteins: five of six K-turns in 23S rRNA make significant interactions with at least one ribosomal protein.

5.3.8.2 Bracket notation

rRNA secondary structures can be represented by the so-called bracket notation, where a nucleotide sequence is represented by a string of equal length consisting of dots and



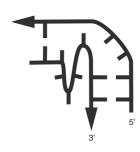


Fig. 5.41 ■ *Left*: Schematic picture of the K-turn found in helix 7 of *H. marismortui* 23S rRNA. *Right*: Secondary structure diagram of a K-turn.

matching parentheses. Unpaired bases are represented by a dot (sometimes a colon), and a base pair between bases i and j is represented by a "('at position i and a')" at position j. Thus, the secondary structure corresponds to the stem-loop structure in Figure 5.42.

```
(((...((((....))))..)))
```

The bracket notation is most easily read from a loop somewhere near the middle of the string. In the example above, the stem-pentaloop structure, with four base pairs in the stem, is given by the construct ((((.....)))), and from there one can elongate the structure in both directions. The elements of the interior loop, with three bases in one strand and two in the other are far apart from each other in the notation. The bracket notation is quite easily understood for small sequences, but for large structures, the notation becomes complex and is mostly suited for input into computer programs.

5.3.8.3 rRNA pseudoknots

A pseudoknot is an rRNA secondary structure containing two stem-loop structures in which the first stem's loop forms part of the second stem. The pseudoknot was first recognized in turnip yellow mosaic virus in 1982, but has since been observed in numerous rRNA structures, the majority from viruses. Pseudoknots fold into a compact three-dimensional structure but are not true topological knots. In fact, a pseudoknot is more of a tertiary-type interaction that forms a double helix.

The two interacting stem-loop regions in pseudoknots do not follow the hierarchical nesting of stem-loop regions required by most secondary structure prediction algorithms. With a slight modification, however, the bracket notation can be extended to also describe pseudoknots. The pseudoknot in Figure 5.43 can readily be described if parentheses are used for one stem-loop structure and square brackets for the other. The resulting notation thus becomes:

Figure 5.44 shows the structure of a pseudoknot-containing aptamer³ isolated from a pool of random sequence molecules. This aptamer was selected for its ability to bind biotin.



Fig. 5.42 ■ The stem-loop structure corresponding to the bracket notation in the text.

³Aptamers are nucleic acid species evolved through repeated cycles of *in vitro* selection to bind to various molecular targets such as small molecules, proteins or nucleic acids.

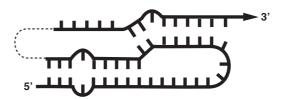


Fig. 5.43 ■ Schematic illustration of a pseudoknot. The pseudoknot consists of two interacting stem-loop structures, so that the loop of one stem-loop is the stem of the other.

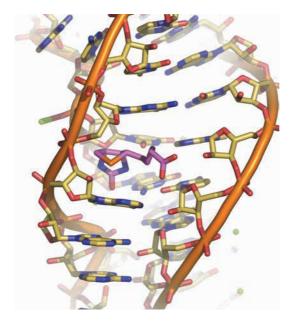


Fig. 5.44 ■ Detail of the biotin binding site in a pseudoknot. Biotin is shown in purple (PDB: 1F27).

The biotin cofactor is bound in a pocket at the interface between the stacked helices of the pseudoknot. Biotin-binding proteins, such as avidin, bury biotin deep in the hydrophobic core, but the aptamer relies on hydrated magnesium ions and water molecules to bind its ligand. This strategy demonstrates the fundamentally different approaches to molecular recognition by proteins and rRNA.

5.3.8.4 The A-minor motif

Base pair interactions between remote nucleotides contribute significantly to the stability of rRNA tertiary structure. It is well known that adenosine is the nucleotide most commonly found in conserved positions outside regular helices. Examination of rRNA in the 50S ribosomal subunit revealed that most of these Adenine bases are involved in tertiary

structure interactions and that these interactions take place according to a limited pattern called the *A-minor motif* (Figure 5.45). Indeed, 26% of the adenine bases in *H. marismortui* 23S rRNA (including 64% of those that are more than 95% conserved) interact with rRNA shallow (minor) grooves via their N1–C2–N3 edges, which are smooth because they lack the exocyclic atoms of the other bases.

The A-minor motif is the predominant tertiary contact in the packing of double helices in the rRNA structures so far studied, so the A-minor motif is perhaps the most important structural element in the formation of rRNA tertiary structure. There are four variants of the A-minor motif that differ with respect to the position of the O2′ and N3 atoms of the A base relative to the O2′ atoms of the base pair in the receptor helix (Figure 5.46). In the type I motif, both the O2′ and the N3 atoms of the A base are inside the shallow groove of the receptor base pair (Figure 5.46). This arrangement optimizes the fit of the

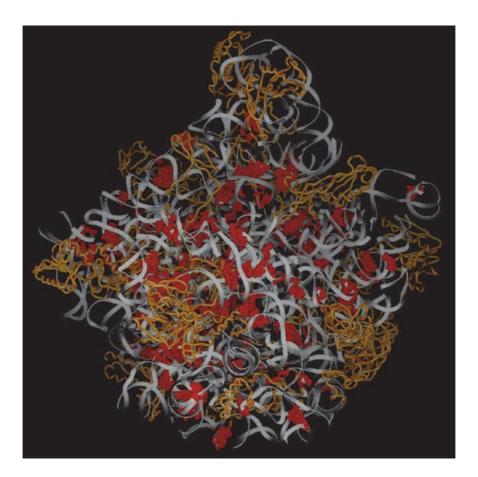


Fig. 5.45 ■ The structure of the 50S ribosomal subunit from *H. marismortui*, showing the 186 adenosines that make A-minor interactions (in red). (Reprinted with permission from Nissen P, Ippolito JA, Ban N, *et al.* (2001) RNA tertiary interactions in the large ribosomal subunit: The A — Minor motif. *Proc Natl Acad Sci USA* **98:** 4899–4903. Copyright (2001) National Academy of Science, USA.)

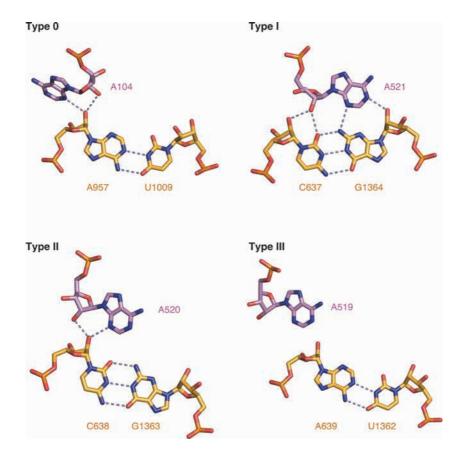


Fig. 5.46 ■ Examples of the four A-minor interaction types from *H. marismortui* 23S rRNA.

adenine into the groove and maximizes the number of hydrogen bonds that can be formed.

In the type II motif, the O2′ of the A base is outside the shallow groove, whereas N3 is inside, and both interact with the 2′-hydroxyl group on the sugar-phosphate backbone of the helix (Figure 5.46). The type III motif is characterized by a positioning of the A base that puts both O2′ and N3 of the A base outside the shallow groove (Figure 5.46). In a fourth, less frequent type 0 motif, the ribose of the A base interacts with a ribose at the receptor helix backbone (Figure 5.46). The type 0 interaction is not base specific because it is the ribose of the inserted residue that fills the shallow groove, and in fact, type III is not base specific either. Nevertheless, in type 0 and type III interactions, the contacts between the inserted base and the receptor helix are optimized when the base is an adenine. The inserted A base of the A-minor motif has a strong preference for GC receptor base pairs, which are optimally complementary in shape and hydrogen bonding pattern.

5.3.9 rRNA is a Carrier of Genetic Information

rRNA has a crucial role as a facsimile or messenger, of genetic information from the chromosomes to the ribosomes. Messenger rRNA (mRNA) is a molecule of rRNA encoding a copy of a single gene or a transcriptional unit. mRNA is first transcribed from a DNA template by rRNA polymerase enzymes (Chapter 10) and carries coding information to the ribosomes where protein synthesis takes place (Chapter 11). In some viruses single- or double-stranded rRNA constitutes the viral genome (Chapter 18).

In the 5' end (the "head" end), the eukaryotic messenger rRNA contains a methylguanosine (m7G) cap, a modified guanine nucleotide that has been added to the mRNA at the time of transcription (Figure 5.47). The presence of the cap is crucial for the recognition of the mRNA by the ribosome, and in addition it protects the mRNA from being degraded by ribonucleases. In the 3' end, the enzyme polyadenylate polymerase adds a sequence of adenine nucleotides to the tail of the pre-mRNA. This poly(A) tail can be up to several hundred nucleotides long. In addition to the coding region, mRNA contains regions that are not translated to protein but have a role in mRNA stability, mRNA localization and translational efficiency. These untranslated regions (UTRs) are present before the start codon (5' UTRs) and after the stop codon (3' UTRs). They contain areas of well-defined secondary structure. Proteins that bind to the UTR regions can interfere with the ribosome's ability to bind mRNA and thus affect translational efficiency. mRNA secondary structure in the 5' UTR region is melted by the initiation factors eIF4A and eIF4B and requires the hydrolysis of ATP. 3' UTRs are believed to be associated with the cellular localization of the mRNA, so that the encoded protein will be translated in the compartment of the cell where it is required. The coding region of mRNA may also contain secondary structural elements, and indeed it is believed that secondary structures conduct translational controls in mRNA.

Thus, mRNA constitutes a comprehensive informational module that not only contains the recipe for constructing the encoded protein but also information about the regulation of the message as well as the destiny of the mRNA molecule itself.

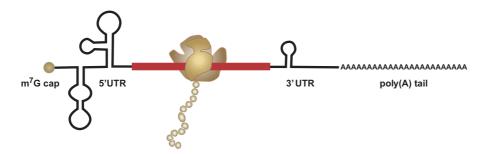


Fig. 5.47 ■ Schematic representation of eukaryotic mRNA showing the 5′ cap, the coding region (red), and the 5′ and 3′ UTRs.

5.3.10 rRNA has Important Structural Roles

The structural role of rRNA is illustrated by ribosomal rRNA (rRNA), which makes up a large part of the ribosomes where proteins are synthesized (Section 11.3). Ribosomal rRNAs are the most conserved genes across the three kingdoms of life, and rRNA sequences are therefore widely used for taxonomic classification and to estimate species divergence.

The 3D structures of ribosomes shows the complex organization of ribosomal rRNA. Originally, rRNA was viewed as no more than a scaffold on which the ribosomal proteins were attached, but this picture is no longer accurate. It is now known that the ribosomal rRNA is closely associated with catalysis. Several of the ribosomal proteins are threaded through the rRNA framework in a manner that precludes an unencumbered folding. Although the smaller rRNAs can form a consistent three-dimensional structure in solution, it has been shown experimentally that the ribosomal proteins participate in the correct folding of the rRNA, and that the rRNA participates in the correct folding of at least some of the ribosomal proteins.

A 17-base sequence loop (15 in prokaryotes), the sarcin-ricin loop (SRL), is one of the most highly conserved in ribosomal RNAs (see Section 11.5.2.1). The fungal toxic enzyme sarcin cleaves a phosphate bond in the loop, and ricin (a toxin from the castor bean) depurinates an important adenine in the motif. The action of sarcin and ricin thus destroys the sarcin-ricin loop, preventing the binding of elongation factors and therefore the proper function of the ribosome.

A structural role can also be ascribed to the UTRs, discussed above (Section 5.3.9). Multiple sequence alignments of UTRs from different mRNAs show a remarkable conservation of secondary structure, which again manifests itself in a conserved threedimensional structure. Thus, the three-dimensional structure of the mRNA has implications for the regulation and control of protein production.

5.3.10.1 *tRNA*

Another class of rRNA that has its primary role in recognition is transfer rRNA (tRNA). tRNA is a class of small RNAs (73-93 nucleotides) active in translation, whose role is to transfer a specific amino acid to the growing polypeptide chain at the ribosomal site of protein synthesis (see Section 11.1.1). Each tRNA species has a three-base anticodon matching a specific codon of the mRNA. In the distal end of the molecule, the tRNA carries the amino acid, covalently attached to its 5' end, matching that specific codon. Each species of tRNA can carry only one type of amino acid, but because the genetic code is degenerate, there are multiple tRNA species per amino acid. The only codons without a matching tRNA species are the stop codons UAG, UAA and UGA. The function of tRNA in translation will be treated in further detail in Chapter 11.

The structure of tRNA was solved in 1974 by two competing laboratories at the MRC, Cambridge, UK and at MIT, Cambridge, USA and it was in fact the first crystallographic structure of a nucleic acid duplex since Watson and Crick put forward their model for the DNA structure.

The sequences of the different tRNA species are very similar, and the 3' end that carries the amino acid is always a CCA single-stranded overhang. tRNA contains many modified bases, especially adenosine when it appears in the first position of the anticodon is always modified to inosine (I), which lacks the amino group on the purine ring. Inosine is lenient in its requirement for base-pairing partners and can base pair with A, U or C, accounting for much of the degeneracy of the genetic code.

The secondary structure of tRNA is a four-stem junction but, due to the characteristic appearance of the secondary structure, is most commonly described as a "cloverleaf" (Figure 5.48). Two of the loops are named after the modified bases they contain, namely, the D-loop (for dihydro-U) and the TYC loop (or just T-loop). The three-dimensional structure of the central, non-helical region of the tRNA junction is very complex. The backbone, the bases, the 2′-hydroxyl groups of the ribose rings, and even water molecules and magnesium ions participate in an elaborate network of interactions that keep the tRNA compact. Figure 5.49 illustrates how the cloverleaf secondary structure folds into a three-dimensional shape. The acceptor stem stacks onto the TYC stem to form a coaxial helix, and likewise, the anticodon stem stacks on top of the linker sequence between the variable loop and the D-stem to form another nearly perfect, coaxial helix.

Tertiary interactions between the $T\Psi C$ and D-loops fold the cloverleaf structure into the L-shaped tertiary structure. Nucleotides G57 and A58 from the T-loop form a stack

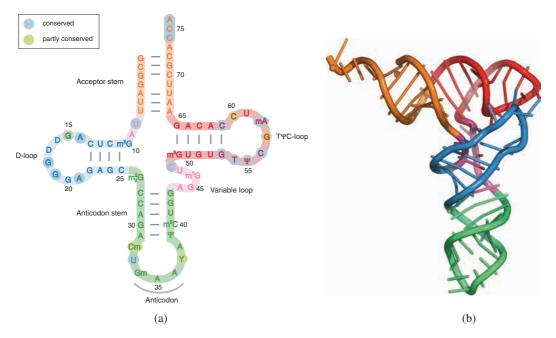


Fig. 5.48 ■ (a) Secondary structure of tRNA^{Phe} from yeast. (b) Schematic representation of the three-dimensional folding of the tRNA molecule, using the same color scheme as in (a).

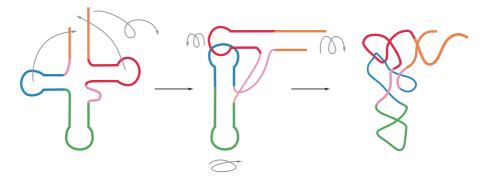


Fig. 5.49 ■ Schematic representation of the folding of the tRNA chain. Many of the conserved residues are responsible for the tertiary structure interactions of tRNA. The coloring scheme is the same as in Figure 5.48.

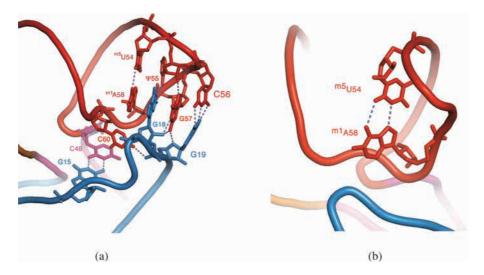


Fig. 5.50 ■ The interactions between the D- and T-loops. Bases from the T-stem intercalate with bases from the D-stem creating a rigid stack. The color code is similar to Figure 5.49.

with G18 and G19 from the D-loop, C56 base pairs with G19, and G15 forms a base pair with C48 of the variable loop. These interactions effectively tie the two parts of the molecule together (Figure 5.50a).

The T-loop wraps around the acceptor stem, and in order to maintain this structure, a reverse Hoogsteen base pair forms across the loop, between positions U54 and m¹ A-58 (Figure 5.50). Mutational studies have shown that while it is possible to mutate these to the isosteric UA pair, it is also possible to have G54:A58 or G54:G58 and maintain the three-dimensional structure. The explanation is that these purine-purine base pairs very closely mimic the reverse Hoogsteen base pair UA and can therefore replace it in the T-loop of a functional tRNA. Other mutations, however, that do not preserve the structural aspects of the 54:58 base pair, fail to form a stable three-dimensional structure.

There are four very sharp turns of the tRNA backbone, one in loop D, one in the variable loop, one in the anticodon loop and one in the T-loop. The latter two are of the classical U-turn motif, characterized by a stabilizing hydrogen bond between a conserved uridine residue and the phosphate backbone further along the chain.

In the anticodon loop of tRNA^{Phe}, nucleotides 34, 35 and 36 compose the anticodon triplet GAA. Positions 33–35 form the U-turn, where the uridine endocyclic N3 forms a hydrogen bond to the phosphate oxygen of A36, and the O2′-hydroxyl of the U33 ribose forms a hydrogen bond to N7 of A35 (Figure 5.51a). This U-turn causes the bases in the anticodon triplet to point directly into solution, poised to interact with the cognate codon.

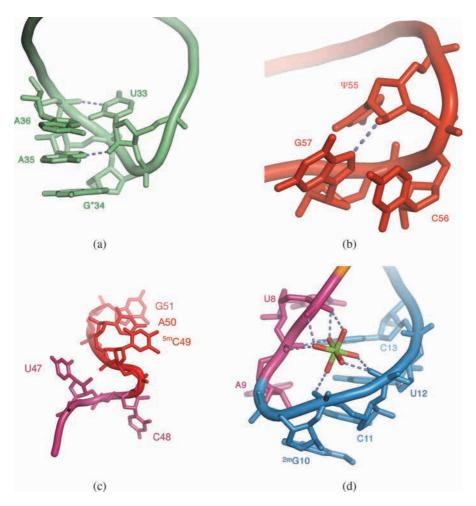


Fig. 5.51 ■ Sharp turns in the tRNA structure. (a) Anticodon loop U-turn 33–36. The turn is stabilized by hydrogen bonds from U33. The anticodon bases extend towards the viewer. (b) T-loop U-turn. (c) Sharp turn of the variable loop. Bases 49–51 form Watson-Crick base pairs with nucleotides in the T-stem and U47 is a part of the D-stem. (d) The sharp turn 9–11, preceding the D-loop, is stabilized by a magnesium ion.

The U-turn motif is also found in the T-loop of tRNA. Residues 55–57 ΨCG form the U-turn here, and the sharp turn of the backbone is produced by the hydrogen bond between N3 of \(\Psi \)55 and the phosphate of nucleotide 58 (Figure 5.51b). In both of these U-turns, the uridine base is positioned at the inside of the loop to interact with the backbone. The sharpest turn of the backbone in tRNA is in the variable loop just between nucleotides C48 and m⁵C49. Interactions with other parts of the molecule effectively pull the bases of these two nucleotides in opposite directions, generating a very sharp turn in the main-chain (Figure 5.51b). This sharp turn helps stabilize the turn in positions 9–11 via a stacking interaction between m⁵C49 and U7. The 9–11 turn is in the linker leading to the D-loop, and it is important because it allows the main chain to reverse its direction from the acceptor stem and fold the "leaf" of the D-loop against the torso of the structure (Figure 5.48). The 9–11 turn assumes its conformation with the assistance of a magnesium ion, which coordinates to the phosphates of the neighboring nucleotides 8, 12 and 13. Finally, the 9–11 turn is stabilized by a reverse Hoogsteen pair between the invariant U8 on the D-stem and the invariant A14 in the D-loop.

A9 is also involved in another type of tertiary interaction: it is intercalated between bases G45 and m'G46. In order to make room between these two bases, the backbone is lengthened by a C2'-endo sugar pucker at m⁷G46.

It is clear that the U-turns and turns of the tRNA structure — which are outside the regular helical regions — are profoundly involved in tertiary interactions and are instrumental in obtaining the correct folding of the molecule.

The core of the tRNA structure also displays several base triples that are involved in maintaining the densely packed core of the molecule. They are situated at the very heart of the structure, in the corner of the molecule's L-shape. The details are shown in Table 5.4 and in Figure 5.52.

Polyamines like spermine and spermidines also play important roles in stabilizing the tertiary conformation of tRNA, and spermine is routinely used in the crystallization of tRNAs and other RNAs. In tRNA, there are two spermidine-binding sites: the first is at the deep groove formed by the T-stem and the acceptor stem. The D-stem and the anticodon arm form the second site. Water molecules and divalent cations also play an

TABLE 5.4 Details of the Base Triples in tRNA

| Triplet | Description |
|----------------------------|---|
| m ² G10:C25/G45 | G45 forms a Sugar/Hoogsteen type of interaction with the major groove of the first Watson–Crick base pair in the D-stem. |
| U12:A23/A9 | A9 hydrogen bonds in the major groove of the U12:A23, forming a reverse Hoogsteen base pair with A23. This stabilizes a sharp turn between bases 9 and 10. |
| C13:G22/m ⁷ G46 | m ⁷ G46 from the variable loop forms a Watson–Crick/Hoogsteen type base pairing interaction with G22 of the D-stem, which is base paired to C13. This anchors the variable loop onto the D-stem. |



Fig. 5.52 ■ Base triples of tRNA. Left: m²G10:C25/G45. Middle: U12:A23/A9. Right: C13:G22/ m⁷G46. See Table 5.4 for details.

integral part of the tRNA structure. Some water molecules mediate contacts between bases and are in effect "extending" the reach of the hydrogen bonding potential of the bases. Six hydrated magnesium ions and three hydrated manganese ions are also found in the structure.

Recent studies on the binding of tRNA to elongation factor Tu (EF-Tu) have shown significant variability in the binding affinities for the torsos of the different tRNA species. That variability is essentially balanced by the protein's variable affinity to the attached amino acid.

5.3.11 rRNA Can Have a Catalytic Role — Ribozymes

Before the discovery of the first ribozyme by T.R. Cech and his research group in the early 1980s, it was believed that only proteins could catalyze enzymatic reactions. Since then, numerous examples of RNAs that catalyze chemical reactions have emerged supporting the "rRNA world" hypothesis of Woese, Crick and Orgel.

The group of catalytic RNAs includes RNAse P (active in tRNA processing), selfcleaving ribozymes such as hammerhead, hairpin, HDV ribozyme and VS ribozyme.

Group I and II introns in mRNA form another class of catalytic RNAs. A group I intron is a self-splicing intron that requires GTP for splicing. It contains an active site that enables it to cleave itself from a precursor mRNA and subsequently to ligate its neighboring exons. This self-splicing reaction can generate an intact mRNA, tRNA or rRNA. Group II introns are found in rRNA, tRNA, mRNA of organelles in certain eukaryotes and in bacterial mRNA. The peptidyl transferase center of ribosomes is also primarily composed of rRNA. The L1 ligase is a man-made rRNA molecule that can catalyze addition of a 5'-triphosphate nucleotide to the 3' end of an rRNA strand. This provides support for an "rRNA world" where ribozymes could accomplish the necessary tasks of self-replication.

5.3.11.1 Hammerhead ribozyme

The story of the hammerhead ribozyme is interesting because it illustrates several points about rRNA tertiary structure. The hammerhead ribozyme catalyzes a transesterification reaction in which the 3′,5′-phosphodiester bond between nucleotides 17 and 1.1 is cut (Figure 5.53). The first structure of hammerhead ribozyme determined in 1994 was an rRNA-DNA hybrid in which a substrate strand was replaced by a DNA strand. This hybrid ribozyme does not have any catalytic activity because the deoxyribonucleotide at position 17 lacks 2′-hydroxyl, which is crucially involved in the cleavage reaction. If nucleotide 17 is replaced with a single ribonucleotide, catalysis can take place.

Another hammerhead ribozyme structure was an all rRNA structure, where the 2'-hydroxyl of cytosine 17 at the cleavage site was replaced with a 2'-methoxyl group in an otherwise unaltered substrate rRNA. The hammerhead was a minimal *in vitro* construct with retained activity, consisting of a 16-nucleotide enzyme strand and a 25-nucleotide substrate strand. This structure was indeed quite similar to the previously determined DNA hybrid ribozyme, except for the presence of some additional hydrogen-bonding interactions. This result was taken as an indication that the crystal structures of ribozyme are closely related to the true solution structure of the unaltered ribozyme.

However, the relationship between the structure of the minimal hammerhead and its catalytic activity became the subject of an escalating controversy. It was simply not possible to reconcile the crystal structure with experiments, and two mutually exclusive hypotheses emerged regarding the nature of the conformational changes required to activate catalysis. One hypothesis required the participation of divalent cations such as Mg²⁺, the other, *acid base catalysis*. In 2006, the crystal structure of a full-length ribozyme from the human parasite *Schistosoma mansoni* was determined, shedding new light on the catalytic mechanism. The three-stem regions align almost perfectly in a continuous

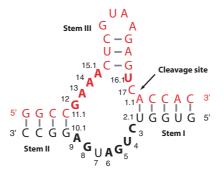


Fig. 5.53 ■ Schematic showing the secondary structure of the minimal hammerhead ribozyme and its similarity to a "hammerhead". The substrate strand is marked in red. Conserved residues in the three-junction central loop are indicated in bold face. The hammerhead nucleotide numbering scheme is also shown.

coaxial helix. Part of stem I forms a side branch to this composite helix, giving the structure the shape of an Icelandic letter *thorn* (Þ) (Figure 5.54). Based on the length of the linker regions, the three-stem junction of the ribozyme belongs to family A. Stem I has a bulge that interacts with the loop of stem II, both through stacking interactions and hydrogen bond formation. This tertiary interaction induces a significant reorganization of the catalytic site compared to the minimal ribozyme structure. A sharp bend is introduced in the backbone region right around the scissile bond at C17.

The bases in the central loop of the junction are closely associated with the active site, although it might not seem so from the secondary structure diagram (Figure 5.54). Residues C3, U4 and G5 form a sharp U-turn, and these bases are oriented towards the exposed ribose of C17, where they assist in positioning that nucleotide. The endocyclic N1 of G12 is within hydrogen-bonding distance of the 2'-hydroxyl group of C17, which is the attacking nucleophile of the reaction. The 2'-hydroxyl group of G8 forms a hydrogen bond to the 5' oxygen of the scissile phosphate, suggesting a possible involvement in stabilization of the leaving group. The catalytic mechanism for hammerhead ribozyme is shown in Figure 5.55. For the reaction to start, the 2'-OH of C17 must be activated, and this is the role of G12, whose endocyclic N1 is close by. However, in its normal protonation state, guanine has a hydrogen atom on N1, but its enol tautomer does not. Assuming that N1 can activate the 2'-hydroxyl of C17, that group can perform a nucleophilic attack on the scissile phosphate attached to the 3' hydroxyl of the same base. The reaction yields a cyclic 2',3'-phosphodiester on nucleotide C17 and a free 5'-hydroxyl on nucleotide 1.1. The mode of catalysis is an in-line mechanism, with the attacking nucleophile, the phosphorous atom and the leaving group on a line. The configuration

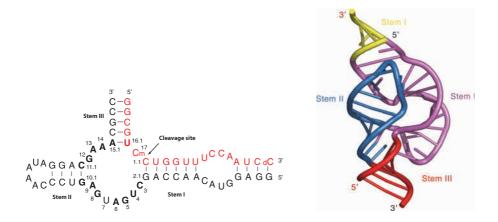


Fig. 5.54 ■ *Top*: Secondary structure diagram of full-length ribozyme (PDB: 2GOZ). Nucleotides in red belong to the substrate strand. Bold face letters indicate conserved sequence motif. Lowercase "d" at the 3' end of the substrate indicates a deoxyribose nucleotide was used to increase synthetic yields. The active site is labeled Cm to indicate that it is 2'-methoxylated. *Bottom*: Schematic figure of the full-length ribozyme. Notice that three stems form a coaxial helix in red, blue and yellow.

Fig. 5.55 ■ Proposed catalytic mechanism for hammerhead ribozyme.

of the reaction intermediate is a trigonal bipyramid in which the apical positions are occupied by the leaving group (5'-OH) and the attacking group (2'-OH).

Figure 5.56 shows the positioning of the important bases around the active site. In the crystal structure, the 2'-OH group of C17 is methoxylated, so the reaction cannot take place. Nucleotide A9 firmly orients G12 with its N1 group poised for activation. The network of hydrogen bonds and stacking interactions introduce a sharp bend in the backbone at the scissile phosphate. This bend in the backbone is also seen in other structures, for example, nucleases. One can almost imagine the backbone "snapping" like a stick. Figure 5.56 also shows a chemical structure diagram of the putative transition state of the reaction, and the involvement of bases G12 and G8 in stabilizing the transition state.

The structures of the minimal ribozymes were not wrong. The crystallographic analyses were completely correct and at a reasonable resolution. Nevertheless, the assumption that the minimal ribozyme structure was identical to the full-length ribozyme surprisingly turned out to be incorrect. Thirty years of experience with protein crystallography has taught us that protein subunits rarely change structures when excised from a larger structure. rRNA structure is more complex and harder to predict and the difference between the minimal and full-length structures is subtle but significant. While the overall folds of the three-stem junctions are nearly identical, the additional tertiary interactions in the full-length ribozyme induce a significant conformational change in the active site region. Importantly, one base, biochemically proven to be essential for the reaction appears near the active site whereas it was 20 Å away in the minimal ribozyme.

Disagreement between a structure and the biochemical experiments should always be taken seriously. It does not necessarily mean that the structure is incorrect, but perhaps that things are not what they seem and that there is possibly a very interesting story hidden in the data.

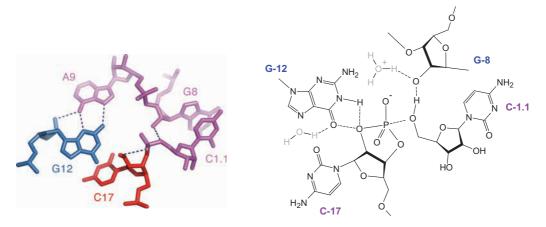


Fig. 5.56 ■ Left: Structural detail of the full-length ribozyme active site. Right: Hypothetical transition state configuration, with G12 initializing the nucleophilic attack and G8 stabilizing the leaving group C1.1.

5.3.11.2 *P4–P6*

The 2.8 Å crystal structure of the 160-nucleotide P4–P6 domain of the *Tetrahymena* group I intron mRNA provides a detailed view of several rRNA folding motifs and exemplifies the fundamental principles of rRNA structure discussed in the previous sections.

Group I self-splicing introns catalyze their own excision from precursor mRNA. This classifies the group I intron as a ribozyme. The catalytic core of the Tetrahymena group I intron consists of two major domains, with the catalytic site split between them. Thus, the P4-P6 domain contains only half of the active site, and therefore has no catalytic activity of its own.

Figure 5.57 schematically shows the organization of P4-P6, consisting of the base paired segments P4, P5 and P6, and the joining regions J3/4, J4/5 and J5/6. P4–P6 also contains an extension element, P5abc (P5a, P5b and P5c), found only in a subclass of group I introns but essential to catalytic activity when it does exist.

Two helical regions, packing side-by-side, dominate the structure of P4-P6 (Figure 5.58). By coaxial stacking, helices P6b, P6a, P6, P4 and P3 form a straight cylinder on one side of the molecule (the conserved core), and the extension element P5abc forms a bananashaped second half.

Two major sets of tertiary interactions form the interface between the extension element and the conserved core. One of these is an interaction between the shallow groove of the P4 helix and the bases in the A-rich bulge. The other is an interaction from the GAAA tetraloop at the tip of the P5abc element to the tetraloop receptor of the conserved core, situated in an internal loop between helices P6a and P6b (J6a/6b).

In the crystal structure, the A-rich bulge forms a turn, with the adenine bases flipped out. The four adenine bases of the bulge are involved in stacking interactions and hydrogen

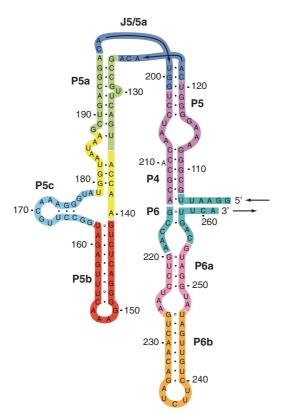


Fig. 5.57 ■ Secondary structure schematic representation of the P4–P6 domain of the group I intron.

bonds, and effectively bridge the two helical stacks of the molecule. To help stabilize the tight turn of the A-rich bulge, two magnesium ions directly coordinate the phosphate oxygens of the loop. The first two adenine bases in the bulge bind to the deep groove of helix P4, whereas the last two bind in pockets at the three-stem junction of the extension element.

The observations on the importance of the A-rich bulge for maintaining the three-dimensional structure of the P4–P6 domain agree well with site-directed mutagenesis and chemical protection studies that highlight the extreme sensitivity of the global structure to mutations at the bulge.

The conformation of the GAAA tetraloop is reminiscent of similar loops observed in NMR studies and in hammerhead ribozyme, confirming that the GNRA fold is indeed a rigid entity. The tetraloop interacts with an 11-nucleotide receptor element in the conserved core of the domain. The three protruding adenosines are in the *anti* conformation and dock into the shallow groove at the P5b and P6a helices, where they form stacking interactions with bases in the receptor element. This stacking is facilitated by adjacent adenosines in the receptor internal loop, which form a specific triple-base interaction, the so-called *A-A platform* (Figure 5.59). In addition to stacking, each adenine of the GAAA loop forms specific hydrogen bonds to the receptor.

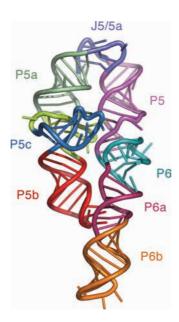


Fig. 5.58 ■ The P4–P6 molecule viewed in the same orientation as the secondary structure shown in Figure 5.57.

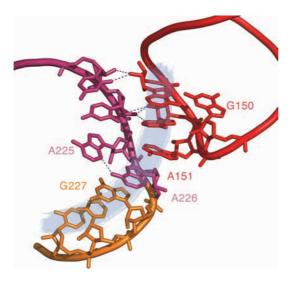


Fig. 5.59 ■ The A-A platform base, consisting of A225 and A226, facilitates the formation of an extended hydrophobic stack, whose interactions are indicated in blue.

The A-rich bulge and the GAAA tetraloop bring the two cylindrical halves of the P4–P6 molecule closely together, resulting in a remarkably close packing of the ribose-phosphate backbone. Two fundamental types of interactions, common in rRNA structures, stabilize this packing.

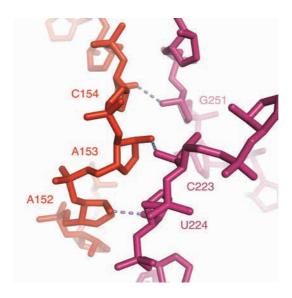


Fig. 5.60 ■ Close-up of the ribose zipper motif from P4–P6, only the sugar-phosphate backbone is shown.

In the first of these, the close packing of phosphates from adjacent helices is mediated by hydrated magnesium ions. In addition, two cobalt atoms from the crystallization buffer are seen snugly fitting into pockets in the structure. Metal ions — magnesium ions in particular — are found in a great number of rRNA structures and are essential for the folding of rRNA into stable tertiary structures and for the catalytic activity of some rRNA enzymes.

The second interaction at the A-rich bulge and in the GAAA tetraloop motifs is the interdigitation of ribose units from the two helical stacks, forming pairwise hydrogen bonds between the 2'-hydroxyl groups. This structural motif is called a ribose zipper and is found in several large rRNA structures, including the 50S ribosome (Figure 5.60).

The P4–P6 domain structure clearly illustrates the importance of A bases in forming the tertiary interactions of even a medium-sized rRNA, and indeed five of the interactions described above are examples of the A-minor motif.

For Further Reading (Sections 5.1 and 5.2)

Review

Dickerson RE. (1989) Definitions and nomenclature of nucleic acid structure components. Nucl Acids Res 17: 1797–1803.

The references to the early work are provided in Chapter 1.

For Further Reading (Section 5.3)

Original Articles

Cate JH *et al.* (1996) Crystal structure of a Group I ribozyme domain: Principles of RNA packing. *Science* **273**: 1678–1685.

Leontis NB, Westhof E. (1998) A common motif organizes the structure of multi-helix loops in 16S and 23S ribosomal RNAs. *J Mol Biol* **283**: 571–583.

Nissen P *et al.* (2001) RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc Natl Acad Sci USA* **98**: 4899–4903.

Reviews

Hendrix DK, Brenner SE, Holbrook SR. (2005) RNA structural motifs: Building blocks of a modular biomolecule. *Quart Rev Biophys* **38**: 221–243.

Leontis NB, Westhof E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA* 7: 499–512.

Lescoute A, Westhof E. (2006) Topology of three-way junctions in folded RNAs. RNA 12: 83–93.

Akiyama BM, Eiler D, Kieft JS. (2016) Structured RNAs that evade or confound exonucleases: Function follows form. *Curr Opin Struct Biol* **36**, 40–47.

Links to Databases

NDB: http://ndbserver.rutgers.edu SCOR: http://scor.berkeley.edu

RNA modification database: http://library.med.utah.edu/RNAmods Non-canonical base pair database: http://prion.bchs.uh.edu/bp_type

Metal binding sites: http://merna.lbl.gov

Basics of Lipids and Membrane Structure

Introduction

More than 100 years ago, E. Charles Overton found that the rates at which nonionic or neutral molecules pass through cell membranes are closely associated with the solubility of these molecules in fluid fat. From experiments on tadpoles, he found that the action of general anesthetics is related to their partition coefficient between water and olive oil. He concluded that membranes are composed largely of lipids, i.e. types of fat-like molecules.

The boundary between a living cell and its surroundings is called the plasma membrane and it is only about 6 nm thick. This membrane barrier, consisting largely of lipids and embedded proteins, controls the flow of nutrients and other molecules into and out of the cell, and responds to hormones and other external signals. Already in 1925, Gorter and Grendel proposed that these membrane lipids are organized into a sheet, two molecules thick, a lipid bilayer as illustrated in Figure 6.1. It shows how the hydrophobic (water-hating), non-polar hydrocarbon chains of the lipid molecules stick together to form a double molecular layer and the hydrophilic (water-loving), polar head groups (the red part of the molecule in the figure) form the interface towards the water solution. Figure 6.1 also shows some other typical aggregate structures that lipids may form.

The curvature of biological membranes can differ within a wide range. Some are very steeply bent such as in mitochondria or the thylakoid of chloroplasts (Figure 6.2). As we will see later in this chapter the lipid building blocks for sharply bent membranes differ from those of flat bilayers.

The most abundant lipids are fats, compounds that are stored by animals and by many plants as an energy reserve. Other lipids form the outer cuticle of plants (complex mixtures of long-chain lipids and hydroxy fatty acids) and yet others serve as protective coatings (often waxes) on feathers and hair. Vitamins A, D, K and E, and ubiquinone are

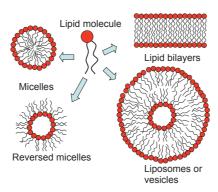


Fig. 6.1 ■ A cartoon of four representative lipid aggregate structures. A lipid bilayer may also form a closed structure called a lipid vesicle or liposome. Note that these drawings only show average geometrical structures. In reality, these structures are much more varied and dynamic.

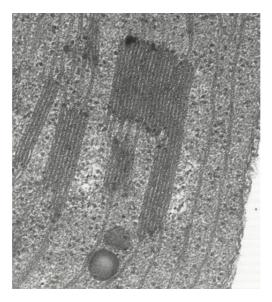


Fig. 6.2 ■ Thylakoid membranes from a chloroplast illustrating the sharp bends (at the arrows) between the flat regions. Electron microscopy picture by C. Weibull and provided by P.Å. Albertsson.

all lipids as are a variety of hormones and light-absorbing plant pigments, such as chlorophylls and carotenoids. These matters and other basic questions are usually discussed at length in common textbooks on biochemistry and biology. In this book, we take a somewhat different approach, usually not seen in ordinary textbooks of biochemistry, and emphasize the structure and dynamics of lipid molecules and their role in the formation of and the physical properties of the membrane. Obviously, lipid bilayers provide the basic structure of almost all biological membranes, but lipids are also crucial for controlling, directly or indirectly, a great variety of biological functions that take place at or are

mediated by membranes. Evidence is also accumulating showing that the functional role of membrane lipids may be as important as that of proteins, and a new science recognizing the role of lipids termed *lipidomics* has emerged in analogy with all the other "omics" in the field of Life Science. A deeper knowledge of the physical properties of lipids is essential for understanding the living world and its modes of functioning. In this chapter, we emphasize the physical properties of the lipid membrane in terms of soft matter with a molecular structured interface. We highlight the phase behavior of membrane lipids, and in particular, several aspects of membrane functioning that involve lipids forming other structures than lamellar bilayers.

This chapter will give the basics for an understanding of:

- why membrane lipids form aggregates,
- when different structures are formed, and
- *how* the physico-chemical properties of the lipids are used by the cell.

6.1 **Molecules That Form Membranes**

6.1.1 Lipid Classes and Their Properties

What is a lipid? In fact, this question has caused some trouble for a long time. A phospholipid is obviously a lipid, but is cholesterol also a lipid? Recently, an attempt to classify lipids has been published (see Further reading). Lipids may be categorized as polyketides, acylglycerols, sphingolipids, prenols or saccharolipids based on their chemically functional backbone. However, for historical and bioinformatics advantages, it was decided to separate fatty acyls from other polyketides, the glycerophospholipids from the other glycerolipids, and sterol lipids from other prenols, resulting in a total of eight primary categories (Table 6.1 and Figure 6.3).

Developing a lipid classification system and nomenclature scheme requires clear guidelines for drawing lipid structures. Several websites provide useful online sources (see Further reading). Here, we provide a simplified version of the lipid structure, sufficient for an understanding of the building blocks of a cell membrane. Unlike proteins, polysaccharides and nucleic acids, most lipids are not polymers, but are made by linking together many smaller molecular parts. Among the "building blocks" of lipids are fatty acids, glycerol, phosphoric acid and sugars. Lipids consist of both polar and non-polar regions, giving them an amphipathic character that accounts for their tendency to aggregate in water into membranous structures and other liquid crystalline phase structures (see Section 6.2.2).

Most membrane lipids have two fatty acyl chains connected to the glycerol backbone. These chains are termed sn-1 and sn-2 in Figures 6.3b and 6.3c. sn for stereochemical

TABLE 6.1 Lipid Categories and Examples. Z Stands for Cis and E for Trans Double Bonds

| Category | Abbreviation | Example | Structure in Figure 6.3 |
|---------------------------|--------------|--|----------------------------|
| Fatty acids | FA | Hexadecanoic acid | (a) |
| Glycerolipids | GL | 1-hexadecanoyl-2-(9Z-octadecanoyl)- sn-glycerol | (b) |
| Glycerophospho- lipids | GP | 1-hexadecanoyl-2-(9Z-octadecanoyl)- sn-glycero-3-phosphatidyl-choline | (c) |
| Sphingolipids | SP | N-(tetradecanoyl)-sphing-4-enine | (d) |
| Sterol lipids | ST | Cholest-5-en-3β-ol | (e) |
| Prenol lipids | PR | 2E, 6E, 10E-farnesol | (f) |
| Saccharolipids | SL | UDP-3-O-(3R-hydroxy-tetradecanoyl)- αd-N-acetylglucosamine | (g) |
| Polyketides | PK | aflatoxin B ₁ | (h) |

numbering is used instead of D and L or R and S. This is the convention used in the lipid area. Note also that although glycerol itself is not chiral, the sn-2 acyl chain is connected to a chiral carbon (Figure 6.3b). Figure 6.4 shows some of the important saturated and unsaturated fatty acids. There is a seemingly endless variety of fatty acids, but only a few of them predominate in a single organism. Most fatty acids contain an even number of carbon atoms due to biosynthetic pathways that add two-carbon units per cycle (see Section 8.4). Plants and animals mainly contain the C_{16} , saturated palmitic, and the C_{18} stearic acid as well as small amounts of the C_{20} , C_{22} and C_{24} acids. Some polyunsaturated fatty acids are essential in the human diet. One of these, arachidonic acid (four double bonds), serves as a precursor for the formation of the hormones prostaglandins and a series of related prostanoids. Our brains contain large amounts of docosahexaenoic (six double bonds) fatty acids in the phospholipids. Bacteria usually lack polyunsaturated fatty acids, but often possess branched fatty acids, cyclopropane-containing acids and hydroxy fatty acids.

The components of complex lipids are linked in many ways, and often glycerol acts as the central unit. Thus, the common fats of adipose tissues and plant oils are triacylg-lycerols or triglycerides, having three fatty acids linked in ester bonds to the glycerol backbone.

6.1.1.1 Phospholipids

As major constituents of biological membranes, phospholipids play a key role in all living cells (exceptions exist — dominant lipids in some bacteria are glucolipids; see

(b) Glycerolipids: 1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycerol

(c) Glycerophospholipids: 1-hexadecanoyl-2-(9Z-octadecenoyl)-sn-glycero-3-phosphocholine

(d) Sphingolipids: N-(tetradecanoyl)-sphing-4-enine

(e) Sterol lipids: cholest-5-en-3β-ol

Fig. 6.3 ■ Representative structures for each lipid category. (Reproduced with permission from Fahy et al. (2005) A comprehensive classification system for lipids. J Lipid Res 46: 839-861. Copyright (2005) American Society for Biochemistry and Molecular Biology.)

$$\begin{array}{c} O_{3} \\ O_{4} \\ \hline \\ O_{2} \\ \hline \\ O_{1} \\ \hline \\ O_{2} \\ \hline \\ O_{2} \\ \hline \\ O_{3} \\ \hline \\ Cg_{2} \\ \hline \\ O_{2} \\ \hline \\ O_{3} \\ \hline \\ Cg_{2} \\ \hline \\ O_{2} \\ \hline \\ O_{3} \\ \hline \\ Cg_{2} \\ \hline \\ O_{2} \\ \hline \\ O_{2} \\ \hline \\ O_{3} \\ \hline \\ Cg_{2} \\ \hline \\ O_{2} \\ \hline \\ O_{3} \\ \hline \\ O_{3} \\ \hline \\ O_{2} \\ \hline \\ O_{3} \\ \hline \\ O_{3} \\ \hline \\ O_{3} \\ \hline \\ O_{4} \\ \hline \\ O_{5} \\ \hline \\$$

Fig. 6.4 ■ Phosphatidylcholine (PC) with some of the most common fatty acyl chains. DPPC stands for dipalmitoyl-PC, POPC for palmitoyloleoyl-PC, PLPC for palmitoyllinoleoyl-PC, PAPC for palmitoylarachidonyl-PC and PDPC for palmitoyldocosahexaenoyl-PC.

Section 6.3.1.2). The two principal groups of phospholipids are the glycerophospholipids that contain glycerol, and the sphingophospholipids that contain the alcohol sphingosine (Figure 6.5). There are a number of different polar head-groups that the phospholipids can contain, for example, choline and ethanolamine that yield zwitterionic head-groups at neutral pH, and serine, glycerol and phosphate that are negatively charged.

Phosphatidylcholines and related phospholipids usually contain a saturated fatty acid in the *sn*-1 position but an unsaturated acid, which may contain between one to six double bonds, at *sn*-2. The *sn*-1 and *sn*-2 carbons of the glycerol backbone are indicated in Figures 6.3 and 6.4. Hydrolysis of the ester linkage at *sn*-2 yields a 1-acyl-3-phosphoglycerol, known as a lysophospholipid. It works like a powerful surfactant or detergent and can cause lysis of cells. Some snake venoms, for example, contain phospholipases that remove one acyl chain on phosphatidylcholine and form lysophosphatidylcholine.

Another group of phospholipids contains hexahydroxycyclohexane or inositol. Phosphatidylinositol is present in the membranes of all eukaryotes and has a specific role in regulating the responses of cells to hormones and other external agents. It also forms part of anchors used to hold certain proteins onto membrane surfaces.

Bacteria and plants often synthesize the anionic phosphatidylglycerol, where the second glycerol is esterified with the phosphate of the polar head-group. Bacteria, as well as mitochondria, contain diphosphatidylglycerol, cardiolipin, in which phosphatidyl groups are attached at both the 1 and 3 positions of glycerol.

In halophilic (salt loving), thermophilic, and methanogenic bacteria, most of the lipids present are phospho- or glyco-lipids (where sugar groups make up the head groups) containing a C_{20} isoprenoid phytanyl group or a C_{40} diphytanyl group, related isoprenoid alcohols or long-chain 1,2-diols (Figure 6.3f). For example, Archaebacteria have a unique set of phospholipids that have ether linkages to their phytanyl chains (instead of ester bonds to acyl chains) as they are derived from archaeol, 2,3-di-O-phytanyl-sn-glycerol. They also

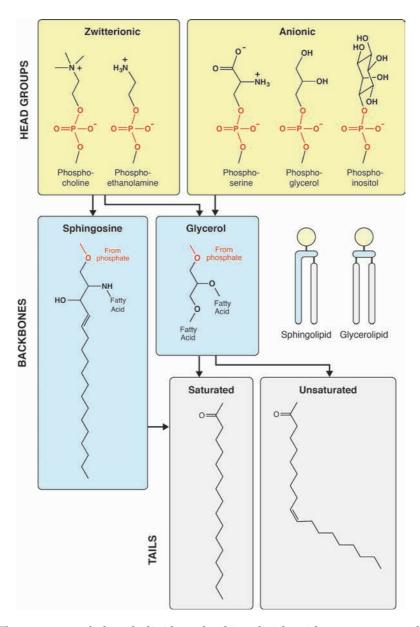


Fig. 6.5 ■ The structure of phospholipids and sphingolipids with some common headgroups.

have unusual head-groups and differ in stereochemistry, esterified to the phosphate head group or sulfated glycolipid on the *sn*-1 instead of the *sn*-3 carbon of the backbone glycerol.

Phosphonolipids that contain a C-P bond are abundant in ciliate protozoa such as Tetrahymena and in some other invertebrates. Phosphonoethylamine replaces phosphoethanolamine in these lipids, which makes the lipids resistant to the enzyme phospholipase C. The phosphonolipids of the outer membrane of Tetrahymena are also ether lipids with an alkoxy group in the sn-1 position. This makes them resistant to phospholipase A_1 as well, and these two properties appear to protect the naked cell membrane of the protozoa from their own phospholipases that are secreted into the environment. Another peculiarity is found in certain marine algae, namely, an arsenic-containing phospholipid O-phosphatidyltrimethylarsonium lactic acid.

6.1.2 Membrane Lipids Form Liquid Crystalline Phases

Since polar lipids are composed of two parts, a hydrophilic part and a hydrophobic part, they are referred to as amphiphiles. Figure 6.6 shows a typical lipid structure, where the two parts in this case are connected by the backbone residue glycerol. In water, such lipids assemble into different types of aggregates that form liquid crystalline phases. It is a well-known fact that oil and water do not mix and this is referred to as the hydrophobic effect. Here, this hydrophobic effect acts to make sure that the fatty hydrocarbon tail of the lipid molecules are screened as much as possible from water and this is achieved by an aggregation of the lipids. Hydrocarbon molecules cannot form hydrogen bonds, and fluid hydrocarbon oil is therefore only held together by dipole-dipole, or van der Waals' interactions. The driving forces that govern the self-assembly of lipids into well-defined aggregate structures are of two kinds: (i) the hydrophobic attraction at the hydrocarbon-water interface, making the molecules associate and (ii) the hydrophilic, ionic or steric repulsion of the lipid head-groups, which imposes the opposite requirement, that they remain in contact with water. These two interactions compete and in 1980, Tanford coined the idea of two "opposing forces", acting mainly in the interfacial region of the aggregate in water. Thus, one force is tending to decrease and the other one

Fig. 6.6 ■ A typical lipid is shown in this figure, here a phospholipid (phosphatidylcholine, PC, often called lecithin). Its amphiphilic character is seen by the hydrophobic hydrocarbon acyl chains (tails) and the hydrophilic polar head group (phosphocholine) connected by the backbone, in this case, glycerol.

tends to increase the interfacial area a per molecule in the aggregate exposed to the aqueous phase.

Lipids and proteins are the major components of cellular membranes and their intimate interplay is increasingly recognized as the basis of the functioning of the biomembrane. In contrast to the long-standing view that lipids only had a subordinate role of an "inert matrix", it is now accepted that lipids are much more essential than just serving as a two-dimensional solvent for proteins. Lipid and water systems exhibit a very rich polymorphism and today some 10 to 15 different phase structures are known. The ability of lipids to form these so-called liquid crystalline phases is also of great interest in a variety of disciplines, above all in colloid chemistry, biological chemistry, material science and structural biology. Liquid crystals are a state of matter that have properties between those of a conventional liquid and those of a solid crystal. For instance, a liquid crystalline phase may flow like a liquid, but its molecules may be oriented in a crystal-like way. Lyotropic liquid crystals exhibit phase transitions as a function of both temperature and composition in the presence of a solvent (typically water), a good example is a lipid/ water system.

Knowledge of the phase diagram of a colloidal or lipid system is a first step in understanding their physico-chemical properties and their possible biochemical role in membranes. Therefore, biophysical chemists spend a lot of time determining phase diagrams of lipids. In general, it is a very tedious, long-term project to determine a complete phase diagram. A mixture of lipid components usually takes a long time (sometimes weeks or longer) to attain equilibrium before one can observe the different liquid crystalline phases formed, since the lipid systems are often very viscous. Only when the phases readily separate macroscopically, it is easy to determine the phase diagram by direct visual observation with the naked eye or in a polarization microscope. It is very common that liquid crystalline phases, even in equilibrium, are dispersed in one another with domain sizes between 10 and several 100 nanometers. Thus, for such systems, visual observation is not sufficient and other methods must be used. The three most common approaches are calorimetry, X-ray scattering and spectroscopic methods like NMR, ESR and fluorescence. Calorimetric measurements are useful only for making a temperature-composition (T-X) diagram. Differential scanning calorimetry (DSC) readily measures the heat capacity as a function of temperature. Small angle X-ray scattering (SAXS) is used for the determination of the gross structure of the aggregates building up the phase. The spectroscopic methods rely on the fact that the properties measured are often sensitive to the local molecular environment. Thus, a spectroscopic investigation of the same molecule in different phases will give rise to different responses. In a multiphase sample, this behavior can be very useful, making it possible to count the number of phases and most often also reveal something about the physico-chemical properties of the phase. In particular, ²H and ³¹PNMR have proven to be very effective (see Further Reading).

Interesting technological applications of lipids are the possibility to polymerize these self-assembling materials in different structures or make use of such lipid microstructures as templates to create stable zeolite-like materials for use in composites, and recently, lipid liquid crystalline phases are used as biocompatible media for e.g. drug delivery. The explosive interest in this aspect of lipids is evidenced by the emergence of commercial enterprises aiming to capitalize on the pharmaceutical and cosmetic applicability of these most versatile materials.

6.1.2.1 How to read a phase diagram

Lipids dispersed in water demonstrate both thermotropic (i.e. a temperature change can induce a phase transition) and lyotropic (i.e. as a solvent can induce a phase transition) polymorphism. The most important thermodynamic principle used to describe the phase equilibria is the Gibbs phase rule that states:

$$F + p = c + 2. \tag{1}$$

F is the number of degrees of freedom that represent the number of independent intensive variables that remain after we have taken all possible constraints into account. (i.e. intensive variable is a variable that is independent of the size of the system (amount of substance), as for example, temperature, pressure or molar fraction will not change if you make the system twice as large). p is the number of coexisting phases in equilibrium and c is the number of components. At constant pressure, F = c - p + 1. For a binary system (exemplified by the dipalmitoylphosphatidylcholine (DPPC)/water system in Figure 6.7) composed of one kind of lipid and water, c = 2 (the minimum number of independent species necessary to define the composition of all the phases present in the system), F = 3 - p.

Frequently, a membrane lipid forms a lamellar (L α) phase (see also Table 6.2) over the whole concentration region (i.e. p=1), and F=2. Thus, both the temperature, and the lipid and water content can be varied in this one-phase region. However, in excess water, where the L α phase is in equilibrium with "pure" water, i.e. p=2, only the temperature can be varied since F=1. At the gel (a kind of a crystalline state) to L α phase transition in excess water p=3, F=0, and the system is invariant. Thus, the gel and L α phases can coexist only at a fixed temperature, often called the main transition, $T_{\rm m}$. Note that for a substance-containing membrane lipids with acyl chains of varying length and unsaturation, a true binary system is not constituted since there may be a large number of different components. However, in general, this does not create any difficulty in the construction of the phase diagram, where these slightly different lipids are all considered as being one lipid, but we should always be aware that for such a "pseudobinary" system "unexpected" results could occur.

For three components (e.g. three different lipids) at constant pressure we have F = 4 - p and it is necessary to also fix the temperature to be able to illustrate the phase diagram in two dimensions. Therefore, for a three-component system we utilize a triangular diagram with the pure compounds in the corners of the triangle (Figure 6.8a,b). The maximum number of phases in equilibrium is three, and a typical characteristic of the

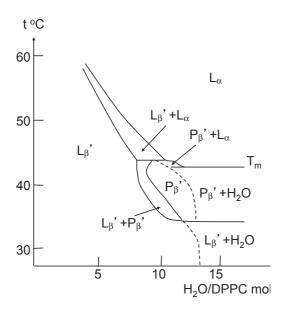


Fig. 6.7 \blacksquare A partial phase diagram of DPPC and water. At low temperature, the gel, L_{β}' , phase is formed and at high temperature and relatively high water content a lamellar liquid crystalline, Lα, phase is stable. In the middle of the phase diagram, the ripple P_{β} phase is stable in a narrow region of temperature and water content. (Adapted with permission from Ulmius J, Wennerström H, Lindblom G, Arvidson G. (1977) Deuteron NMR studies of phase equilibria in a lecithin-water system. Biochemistry 16: 5742–5745. Copyright (1977) American Chemical Society.)

TABLE 6.2 Liquid Crystalline and Gel Phases Formed by Membrane Lipids

| Phase | Dimension | Nomenclature | Order |
|-------------------------------|--------------------------|-------------------|------------------------------------|
| Lamellar | One | L_{α} | Disordered, fluid |
| Ripple gel | Two; oblique or centered | $P_{m{eta}'}$ | Rippled |
| Gel | One | $L_{\pmb{\beta}}$ | All-trans ^a acyl chains |
| Normal hexagonal | Two | $H_{\rm I}$ | Disordered, fluid oil-in-water |
| Reversed (inverted) hexagonal | Two | H_{II} | Disordered, fluid water-in-oil |
| Cubic | Three | I | Disordered, fluid |
| Normal cubic | Three | ${ m I_I}$ | Disordered, fluid |
| Reversed cubic | Three | ${ m I_{II}}$ | Disordered, fluid |

^a Here all-trans refers to an acyl chain conformation (see any textbook on organic chemistry).

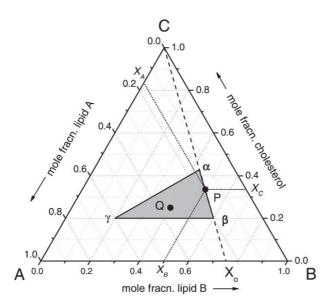


Fig. 6.8a • Representation of the compositions and phases at constant temperature and pressure for a ternary phase diagram of lipids A, B and cholesterol, C. The mol fractions, X_A , X_B and X_C are indicated along the edges of the equilateral triangle, and the corners of the triangle represent 100% pure substance A, B or C. α –β, β – γ and α – γ are tie lines and α β γ is a three-phase triangle (three phases in equilibrium). (Adapted with permission from Marsh D. (2009) Cholesterol-induced fluid membranes domains: A compendium of lipid-raft ternary phase diagrams. Biochim Biophys Acta 1788: 2114–2123. Copyright (2009) Elsevier.)

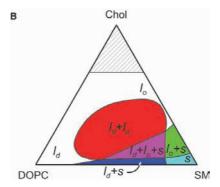


Fig. 6.8b ■ Rough estimate for the boundaries of one-, two- and three-phase regions in the phase diagram of the ternary mixture DOPC/SM/Chol at 23°C, where SM is sphingomyelin extracted from egg, DOPC denotes dioleoylphosphatidylcholine and Chol is cholesterol. The different onephase regions are denoted by l_d , l_o and s (the structures of these three different phases are shown in Section 6.3.6, where lipid domains are discussed). The number of phases are indicated in the various two- and three-phase regions by $l_d + l_o$, $l_d + s$, $l_o + s$ and $l_d + l_o + s$, respectively. The phases in the hatched area at the top of the diagram consists of l_d and cholesterol monohydrate crystals. (Reproduced with permission from Bezlyepkina N, Gracia RS, Shchelokovskyy P, Lipowsky R, Dimova R. (2013) Phase diagram and tie-line determination for the ternary mixture DOPC/eSM/ cholesterol. Biophys J 104: 1456–1464. Copyright (2013) Biophysical Society.)

ternary phase diagram is the areas of the three-phase triangles. This is in comparison to the three-phase lines present in the two-component systems. The compositions of ternary mixtures (at the point P in Figure 6.8a) in mole fractions (or sometimes in wt%) are given by the set of triangular coordinates $(X_A, X_B \text{ and } X_C; X_A + X_B + X_C = 1)$ as indicated in the figure. Note that addition of cholesterol (one of the three components in the example in Figure 6.8a), C, to a mixture at point P occurs along the hatched line X_oC, where the mole ratio between the lipids A and B (X_A/X_B) are constant.

It turns out that ternary phase diagrams play an important role in the formation of rafts, as discussed in Section 6.3.6, and in particular phase separation in bilayers from ternary mixtures of cholesterol with two lipids that have a low and a high chain-melting temperature, respectively. An example is shown in Figure 6.8b. The details of the phases formed are discussed in 6.4.5.

In the construction of phase diagrams the so-called *lever rule* is very useful. A point in a two-phase region of a phase diagram (binary or ternary) indicates not only qualitatively that two phases are present but represents quantitatively the relative amounts of each one. The relative amounts of the two phases that are in equilibrium are determined by the relative distances of the particular point on its tie line (a line joining two points representing phases in equilibrium) from the respective phase boundaries — this is called the lever rule. For a binary system, tie lines are always horizontal, but for a ternary system their directions are not always easily predicted, and they have to be determined experimentally. Here, the NMR method is particularly convenient, since the area under a peak in the NMR spectrum is proportional to the number or fraction of nuclei giving rise to the signal. This can be used to determine the proportions of different phases in the sample under study.

The following is the nomenclature most commonly used when describing the different phases formed by lipids. The upper-case Latin letter characterizes the type of long range order (one-, two-, or three-dimensional lattice), the subscript Greek letter stands for ordered (β) or disordered (α) acyl chains, and the subscripts I and II denote normal and reversed liquid crystalline structures, respectively.

The following *liquid crystalline phases* and so-called gel phases are the most frequently occurring for membrane lipids (Table 6.2 and Figure 6.9). Note that for the gel phase the acyl chains are in a crystalline all-trans state.

An example of a binary phase diagram containing several different lamellar and non-lamellar phases is shown in Figure 6.10. In the phase diagrams in Figures 6.7 and 6.10, it can be seen that by lowering the water content, phase changes are triggered between liquid crystal and gel or between different liquid crystalline phases. This is due to the fact that phospholipid bilayers interact through a short-range repulsive force, and the lower the water content, the stronger this repulsive interaction becomes. Such an interaction between aggregates (interbilayer interaction) may cause phase changes. Thus, there is a close relationship between forces, between the bilayers and the phase behavior. With decreasing water content, the bilayer is more likely to transform to a condition in which the repulsive interaction is weaker. At the phase transition, the interaction between the bilayers compensates for the difference in the interactions within the bilayers (intrabilayer interaction).

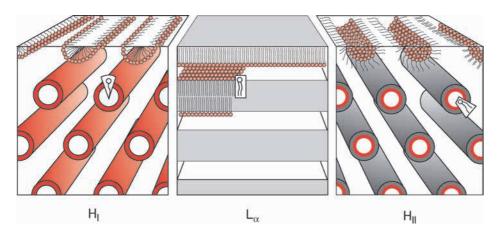


Fig. 6.9 ■ Structures of some common liquid crystalline phases. From *left* to *right* (i.e. with decreasing water content): normal hexagonal phase, H_{II} , lamellar phase, L_{α} and the reversed (inverted) hexagonal phase H_{II} . Note the geometrical shape of the lipids indicated in the different phase structures (see Section 6.2). (Adapted with permission from Lindblom G, Rilfors L. (1989) Cubic phases and isotropic structures formed by membrane lipids — Possible biological relevance. *Biochim Biophys Acta* 988: 221–256. Copyright (1989) Elsevier.)

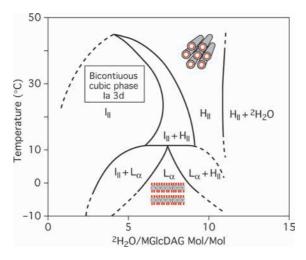


Fig. 6.10 ■ A partial phase diagram of the system monoglucosyldiacylglycerol (MGlcDAG)/heavy water, where MGlcDAG comes from the bacterium *A. laidlawii*. Note that an L_{α} phase is stable only at a temperature that is much lower than the growth temperature (37°C), i.e. only the non-lamellar phases H_{II} and I_{II} are formed at growth temperatures (see Section 6.3.1.2). Adapted with permission from Lindblom G, Rilfors L. (1989) Cubic phases and isotropic structures formed by membrane lipids — Possible biological relevance. *Biochim Biophys Acta* 988: 221–256. Copyright (1989) Elsevier.)

6.1.2.2 The skin membrane — application of a lipid phase diagram

The skin is our largest organ and the main permeability barrier of the skin is the outermost horny part, the so-called *stratum corneum* (SC). This barrier is a thin (ca. 20 µm) dry layer that separates very different environments, namely, the water-rich inside of the body from the dry outside. Moreover, it can be exposed to rather extreme gradients in water and temperature, and also to chemicals such as drugs. The structure of SC is built up like a brick wall, where the bricks are dead cells (so-called corneocytes) and the "mortar" between them is made of lipids forming bilayer structures. The unique barrier properties of this remarkable organ are largely determined by these lipids. The permeability may be quite large when these lipids are in the fluid state, while it is about zero when they are in the crystalline form.

The physical environments on the two sides of the human skin are profoundly different. Since lipids can adopt a range of different phase structures depending on the water content and temperature (see phase diagrams in Section 6.1.2.1), the lipids prefer one structural arrangement on the water-rich side of the barrier membrane, while on the other (more dry) side a different structure is preferred. Such a scenario will occur across the (human) skin and therefore, its barrier properties will be affected by changes in the atmosphere outside the skin. Now, the phase structure of the lipids in the "mortar" is determined by the phase diagram of the particular lipids present in the mortar. For example, it can be inferred from the phase diagram in Figure 6.7 that as the water content at constant temperature is reduced for an L α phase there will be a transition from this fluid lamellar phase to a crystalline gel phase (Lβ). Furthermore, it has been shown experimentally that as the water content is increased for an Lα phase (above 10%), the thickness of the lipid bilayers remains constant. However, the distance between the bilayers increases with the water content (this is illustrated in Figure 6.11. The water content is increased from top to bottom in the pictures with red and green frames). As schematically shown in the red frame to the left in Figure 6.11, starting at the top in the figure with a solid gel phase (Lβ), an increase in the water content will not only increase the distance between bilayers but also result in a phase transition to an La phase. Remember that the crystalline, solid phase (LB) is obtained at low relative humidity (RH, low water content in Figure 6.7). Thus, variations in the structure within the lipid membrane can lead to rather dramatic changes in the membrane barrier properties, and this feedback mechanism has been treated in theoretical and experimental studies of model membranes (Figure 6.11). The changes in the lipid phase affect the permeability of the SC depending on the water content or humidity outside the skin. Thus, one important aspect of the SC membrane is that its properties are regulated by changes in its environment. As an example, there is an abrupt increase in skin permeability towards model drugs at high degrees of skin hydration (Figure 6.11), which is also taken advantage of in transdermal delivery applications (called occlusion effect). These observations can likely be explained by molecular

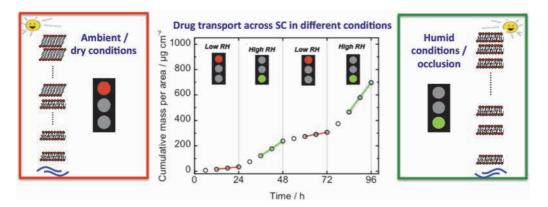


Fig. 6.11 • Molecular explanation for occlusion effect in skin. The permeability of a drug across a stratum corneum (SC) membrane at different humidities is shown together with schematic illustrations of lipid multilamellar membranes in the absence and presence of an osmotic gradient. The osmotic gradient may induce heterogeneous swelling, and phase transitions between fluid and solid lamellar phases, which in turn affect the membrane barrier properties. Note that at high relative humidity (RH) the distance between bilayers increases with increasing water content (from top down; green frame), while at low RH, there is also a phase transition between crystalline, solid gel to fluid lamellar phase with increasing water content (from top down; red frame). In the middle diagram experimental data for the steady-state flux of a model drug (metronidazol) across an intact porcine SC is shown. The membrane barrier can be switched on and off by varying the external water gradient. (Data from Björklund S, Engblom J, Thuresson K, Sparr E. (2010) A water gradient can be used to regulate drug transport across skin. J Contr Rel 143: 191-200. Courtesy of Emma Sparr.)

rearrangement in the SC membrane in response to the changes in the water gradient across the membrane. NMR studies of intact SC have demonstrated that hydration increase fluidity in SC components (see Further Reading).

6.1.2.3 Closed spherical lipid bilayers or vesicles

In isotropic solutions, lipid bilayers can form spherical shells rather than infinite planar bilayers. This occurs because in a closed bilayer the energetically unfavorable edges are eliminated. The size of vesicles can differ significantly, and we distinguish between small unilamellar vesicles (SUV) having a radius <100 nm, large unilamellar vesicles (LUV) and multilamellar liposomes (an L_{α} phase in excess water). The vesicles are seldom true thermodynamically equilibrated systems but are often metastable over a long time (days to weeks) and can be used in various experiments involving membrane proteins, lipid domain formation or vesicle fusion.

The nature of the vesicle solution depends on how it is prepared. A common and simple way of producing vesicles is to subject multilamellar liposomes to ultrasound. The

result of such a treatment is a broad but skewed size distribution of vesicles having rather small aggregates, typically around 150-200 Å. Another way to make vesicles is to use cholate dialysis with a better control of the resulting preparation. These vesicles possess a more uniform size, and can be controlled by parameters such as ionic concentration, temperature and pH. Finally, a final method by which vesicles can be produced is to inject a solution of the lipid dissolved in an organic solvent or solvent mixture into excess water. Large unilamellar vesicles (LUV) or even giant unilamellar vesicles (GUV) form as the organic solvent evaporates or is dissolved in water. The kinetic stability of a vesicular system depends on the rate of the process by which two vesicles fuse to form a larger one, and this is a good example of the general problem of colloidal stability. The fusion process is discussed in Section 6.3.3.

6.2 **Amphiphile Self-Assembly Into Different** Aggregate Structures

6.2.1 Lipid Packing and Spontaneous Curvature

One of the most useful concepts for a qualitative understanding of the phase behavior in amphiphilic systems is based on the geometry or general shape of a lipid molecule (Figure 6.12). The self-assembly of lipid molecules depends on a dimensionless packing parameter defined by the ratio:

$$P = v/al$$
.

where v is the volume of the fluid hydrocarbon chains, l is the molecular length and a is the optimal cross-sectional area¹ of the polar head group as shown in Figure 6.12.

When the packing parameter (sometimes also called the surfactant number) is less than one, the shape of the lipid molecules are cones (Figure 6.12) that can pack themselves into spherical micelles or H_I aggregates (Figure 6.13). When it is equal to unity (P = 1; cylindrical-like molecules, Figure 6.13), the conditions are optimal for the formation of a bilayer structure. If P > 1, the lipid molecules are wedge-shaped and the lipid monolayer prefers to curve towards the water region, i.e. it forms reversed micelles or an $H_{\rm II}$ liquid crystalline phase (Figure 6.13).

¹It should be noted that the cross-sectional area is usually not a quantitative measure of the size of the headgroup, i.e. you cannot measure the area just by looking at the molecule. It depends on many factors in the surroundings, like pH, charge, electrolytes, etc. This is particularly important for ionic amphiphiles. However, we can always predict the trends in the area a for a specific change in the system, and this ability can be extremely useful for qualitative interpretation of experimental data.

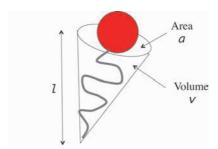


Fig. 6.12 ■ A schematic drawing of the shape of a lipid molecule forming spherical micelles as shown in Figures 6.1 and 6.13. The red sphere is the polar head group and the hydrophobic tail is shown in gray. The components defining the packing parameter are indicated.

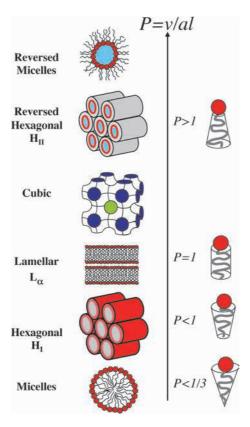


Fig. 6.13 ■ Lipid molecules of different shapes and packing parameters. The possible aggregate structures in different phases (micellar solutions and liquid crystalline phases) are shown to the left. A liquid crystalline phase has the properties of a liquid and at the same time shows a long-range order as in a crystal.

This simple approach is very useful for explaining the kinds of shapes certain molecules assume. However, one has to always remember that the surface area *a* might have a complex dependence on temperature, charge and salt content or pH, and more sophisticated considerations may be needed. For example, a change in the molecular shape does

not fully explain why a reversed hexagonal, H_{II}, phase is formed at high water contents when an alkane or hydrophobic peptide is added to seemingly stable lamellar liquid crystalline phases of phosphatidylcholine (PC), an apparently "cylindrical" lipid molecule. Obviously, this suggests that the PC molecules, when forming a monolayer, have a packing parameter slightly larger than one, but other factors restrain them from forming a curved monolayer. In fact, it is not possible to pack the PC molecules in a large H_{II} cylinder without creating a large interstitial volume of vacuum as will be discussed below.

Bilayers formed by such PC molecules are said to be "frustrated" (see below). This is explained by a concept known as lipid monolayer elasticity related to the packing parameter but with a more general character not involving the lipid molecules specifically. The energy needed to deform a membrane is determined by the structure and elasticity of the membrane. The non-deformed unstressed state of the membrane is referred to as the spontaneous state. Deviations from the spontaneous state, the forces required for these deviations, and the accumulated energy in the new shape determine the membrane's elastic properties.

To understand this, let us briefly review the physical chemistry of membrane bending and the energetics involved. First, we need to look at some definitions. At any point on a sheet in three-dimensional space, two principal radii of curvature R_1 and R_2 and local curvatures $c_1 = 1/R_1$ and $c_2 = 1/R_2$ can be defined (Figure 6.14).

The sign of the curvature is arbitrary, and by convention one uses a definition as shown in Figure 6.14, where a region that bulges "outward" from the volume enclosed from the surrounding medium has a positive curvature. Thus, spherical vesicles have uniformly positive curvature, since R_1 and R_2 are both positive and equal. Saddle-shaped membranes found, for example, on the bicontinuous cubic phase structure (illustrated in Figure 6.16) or at the necks of budding vesicles, have positive curvature along one principal axis and negative along the other. The energetic cost per unit area associated with bending a monolayer is described by the Gibbs elastic curvature energy, and it is given by the sum of two terms, one dependent on the *total curvature* of the monolayer $(c_1 + c_2)$ and the other on the product of c_1 and c_2 :

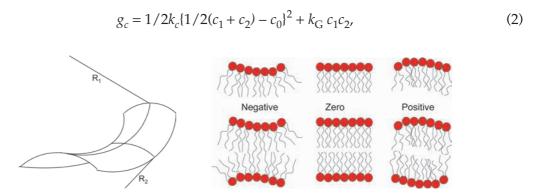


Fig. 6.14 ■ Left: The definition of the two radii of membrane curvature. In this case with a saddleshaped surface the two radii have different signs. Right: Illustration of the definition of the sign of the radius of curvature (by convention).

where k_c is the elastic bending constant and k_G is the saddle-splay (or Gaussian curvature) constant. $c_0 = 1/R_0$ is the *spontaneous curvature*, i.e. the monolayer radius, R_0 , of a relaxed, stress-free state of the lipid monolayer. The spontaneous curvature is a measure of the tendency of lipid monolayers to bend into non-planar geometries. (The total elastic curvature energy, G_c , is obtained by integrating g_c over the area of the monolayer.)

It can thus be seen that it is advantageous to have the mean curvature, $1/2(c_1+c_2)$, close to the spontaneous curvature c_0 to minimize the free energy of curvature. In the structure of an H_{II} phase the lipid monolayers bend to form cylinders of radius R. It turns out that the Gibbs energy, involved in the formation of an $H_{\rm II}$ phase contains two parts. One part is the energy of curvature, and, since the acyl chains of the lipid molecules must stretch to fill the hydrophobic regions between proximate cylinders in the H_{II} phase (the green areas in Figure 6.15), the other part is the non-zero packing energy, g_n . The total Gibbs energy is then the sum of the elastic curvature and packing energy, $g_{tot} = g_c + g_p$. We thus have a situation where two physical forces, curvature and packing, oppose one another, and such a situation is referred to as a "frustration" as mentioned above. It has been shown that this "frustration" can be decreased or eliminated by the addition of hydrophobic molecules, e.g. alkanes. These molecules preferentially partition into the hydrocarbon regions in the middle between the cylinders (the green areas) in an $H_{\rm II}$ phase, and in this way fill the void volumes created in the formation of the H_{II} structure (Figure 6.15).

Finally, it should be noted that transmembrane peptides may in certain situations induce an H_{II} phase, although the mechanism behind the process is different than the effect of alkanes (see Further Reading).

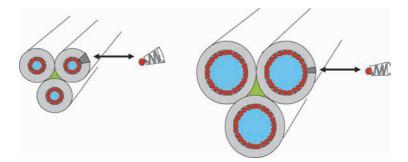


Fig. 6.15 ■ As the lipid molecules (enlarged) get less wedge-shaped the radius of curvature of the cylinders building up the H_{II} phase increases and the water (blue) uptake increases in the H_{II} cylinders. The green void volume (hydrophobic) will also increase and the H_{II} phase will only form if hydrophobic molecules, like alkanes, are present in between the cylinders. Note that a lipid with saturated chains forms larger cylinders in the H_{II} phase than a lipid with unsaturated chains. (Adapted with permission from Sjölund M, Rilfors L, Lindblom G. (1989) Reversed hexagonal phase formation in the lecithin-alkane-water systems with different acyl chain unsaturation and alkane length. Biochemistry 28: 1323-1329. Copyright (1997) American Chemical Society.)

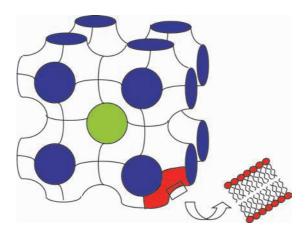


Fig. 6.16 ■ Schematic illustration of the structure of one of the bicontinuous cubic phases (sometimes called the plumber's nightmare). Note that lipid monolayers are draped on the in- and outside of a minimal surface (Schwarz's surface), creating a "wrinkled" lipid bilayer. There are two "water systems"; a water molecule in the blue regions will never pass into the water regions marked green. (Adapted with permission from Lindblom G., Rilfors L. (1989) Cubic phases and isotropic structures formed by membrane lipids — Possible biological relevance. Biochim Biophys Acta 988: 221–256. Copyright (1989) Elsevier.)

Another fascinating liquid crystalline structure with curved monolayers is found in bicontinuous cubic phases. Here, monolayers are draped on each side over a minimal surface (a surface where every point has a curvature equal to zero, i.e. $c_1 + c_2 = 0$ everywhere on the minimal surface). This surface describes the midplane of the bilayers and not the interface between the polar and hydrophobic regions (Figure 6.16). Because of this structure, the "frustration" is usually less for the cubic phase than for the L_{α} and H_{II} structures, and the cubic phase, therefore, frequently appears between the L_{α} and $H_{\rm II}$ phases in lipid phase diagrams. The three-dimensional cubic phase has been identified as a useful medium for the crystallization of membrane proteins.

Consequently, curvature energy plays an important role for the stability of bicontinuous cubic phases. Another liquid crystalline phase, observed especially in many nonionic amphiphile-water systems, is the isotropic L₃ phase (often referred to as the sponge phase) that generally is in equilibrium with both a dilute solution and an L_{α} phase. Usually, the sponge phase has a very narrow range of stability both in temperature and composition. The basic structural unit in the L_3 phase, like in many cubic phases, is a network of connected lipid bilayers. The driving force behind the formation of an L₃ rather than an L_{α} phase is the opportunity to model an optimal curvature of the lipid monolayer. The structure of the L₃ phase arises from a melted or disordered cubic structure (Figure 6.17). Such a disordering is favored by weak interactions between the bilayers. At a high lipid concentration, and thus strong interbilayer forces, a cubic phase may form in equilibrium with the L₃ phase.

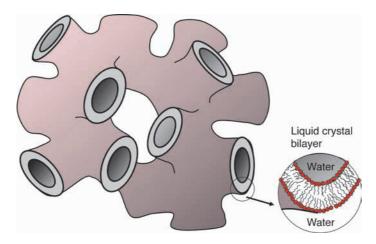


Fig. 6.17 \blacksquare A schematic picture of a sponge (L₃) phase.

6.2.1.1 Lipid packing and lateral pressure

Another important physico-chemical property of the lipid bilayer is lateral pressure. As mentioned above, the hydrophobic effect that keeps the lipid chains away from the water results in the aggregation of lipid molecules to a lipid bilayer. However, the lipid molecules are often subjected to great stresses because they are confined to a bilayer structure where the packing conditions can lead to frustration, as mentioned earlier. This results in one of the most fundamental physical properties of the lipid bilayers, namely, the lateral pressure profile (see Figure 6.18).

As can be inferred from Figure 6.18, there are different forces involved in the stabilization of the bilayer aggregate. For a bilayer in equilibrium the forces, of course, have to cancel and give a zero net force. Since the forces operate in different planes, the pressures are distributed non-evenly across the bilayer. The lateral pressure profile contains three contributions: a positive pressure resulting from the repulsive force between the head groups, a negative pressure because of interfacial tension at the hydrophobic/hydrophilic interface and a positive pressure arising from entropic repulsion between the flexible lipid hydrocarbon chains (compare the discussion in 6.2.1 on the phase transition to an H_{II} phase upon addition of an alkane to a lipid bilayer). Therefore, the lateral pressure profile depends on the lipids building up the bilayer, i.e. lipids with a bulky fatty acid cause greater chain pressure than a saturated straight chain. Since the bilayer is very thin, the large interfacial tensions from the two interfaces have to be distributed over a very short range. This means that the counteracting pressure from the lipid chains has to have a huge density, characteristically several 100 atmospheres. The interfacial tension (γ) at both the bilayer interfaces is about 50 mNm⁻¹. It is then easy to calculate the lateral pressure in the interior of the lipid bilayer, which has to counter-balance this tension

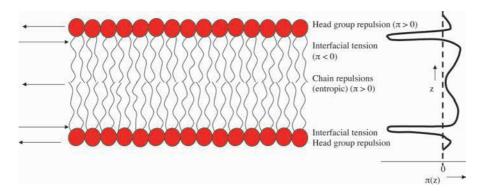


Fig. 6.18 \blacksquare Illustration of the lateral pressure, p(z), profile in a lipid bilayer. A coordinate system, z, along the normal to lipid bilayer, showing the pressure distribution across the bilayer is schematically indicated to the right. The lateral pressure in the middle of the bilayer can be very high. However, the total pressure over the bilayer is zero. (Courtesy of Ole Mouritsen.)

over a distance of the bilayer thickness, d = 2.5-3 nm. The lateral pressure will be equal to $2\gamma/d$ = about 40000 kPa, i.e. almost 400 atm. A lateral pressure of this magnitude can influence membrane proteins, for example, by changing the protein structure (Figure 6.19).

Analogously, the introduction of a lipid of a different shape into the bilayer membrane will change the lateral pressure allowing the cell to open or close a membrane protein channel or pore. Such a channel is called mechanosensitive. These channels act as membrane-embedded mechanoelectrical switches, opening large water-filled pores in response to lipid bilayer deformations. This process is critical to the response of living organisms to direct physical stimulation, such as touch, hearing and osmoregulation. An example is the large prokaryotic mechanosensitive channel (MscL) illustrated in Figure 6.20. The open state is highly dynamic, supporting a water-filled pore of 25Å. The channel was shown to open when lysophosphatidylcholine (LPC) was added to the cell membrane, dramatically lowering the activation threshold, and relieving the lateral pressure on the protein (Figure 6.21).

6.2.1.2 Lipids move fast in the bilayer

The presence of lively dynamics in lipid bilayers accentuates the difficulty in drawing one simple picture or model of a lipid membrane. The useful cartoons of a lipid bilayer shown in Figure 6.1 or the structures in Figures 6.9 and 6.13 should be seen as appearances averaged over a relevant timescale (in the range of milli- to micro-seconds), and these pictures do not provide the full information about possible dynamical aspects of, for example, trans-bilayer structures occurring upon protein transport across membranes.

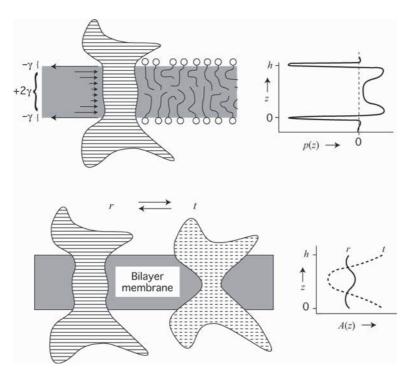


Fig. 6.19 ■ High lateral pressure, p(z), can result in a change in the conformation of an integral membrane protein (striped or dashed) as illustrated by the cross-section A(z). The protein can be in any of two states; r or t. γ is the interfacial tension. (Reprinted with permission from Cantor, RS. Lateral pressures in cell membranes: A mechanism for modulation of protein function. J *Phys Chem B* **101**: 1723–1725. (Copyright (1997) American Chemical Society.)

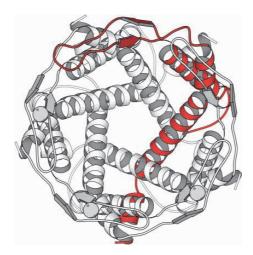


Fig. 6.20 ■ The structure of the closed state of the large membrane channel, MscL, from the *M. tuberculosum* outer membrane viewed perpendicularly to the membrane surface. This is the membrane part formed by five subunits, each contributing two transmembrane helices. One subunit is in red (PDB: 2OAR).

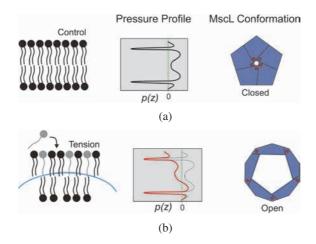


Fig. 6.21 ■ A schematic picture of the effect on the bilayer curvature (and therefore on the lateral pressure) and a mechanosensitive protein (MscL). (a) The resting state with the channel closed. (b) The channel opens when a lipid with a single hydrocarbon chain — like lysophosphatidylcholine — is added to the membrane. (Adapted with permission from Perozo E & Rees DC. (2003) Structure and mechanism in prokaryotic mechanosensitive channels. Curr Opin Struct Biol 13: 432–442. Copyright (2003) Elsevier.)

A comparison of the properties of lipid bilayers and spherical micelles formed by single chain surfactant molecules or lysolipids provide useful insight into distinctive molecular prop erties of bilayers. Thus, the amphiphiles in micellar aggregates exchange with the monomers in the water solution in the range of 10^{-3} to 10^{-6} s⁻¹. In bilayers, the monomer exchange is considerably slower, as can be predicted by their much lower monomer solubility $(10^{-5} \text{ to } 10^{-10}$ M). The aggregate lifetime for typical micelles vary from 10^{-3} to 10^{-1} s, while for bilayers it varies from days to years. Therefore, it may require hours, days or even years for a bilayer to achieve equilibrium state. However, within the bilayer the lipid molecules move very fast and undergo a range of different dynamical processes. They are constantly changing intramolecular conformations, they are wobbling, they are protruding out of the layer, and they are diffusing around laterally. These motions range over an enormous time span, from picoseconds to hours. Conformational changes are fast, since they involve rotations around C-C bonds, which typically take a few picoseconds, and the rotation of the lipid molecules occurs on a time scale of nanoseconds, whereas lateral diffusion is in the range of tens of nanoseconds. A typical lipid will on average rotate once around its axis while it travels a distance corresponding to its own size. The wobbling of the hydrocarbon chains that leads to changes in the orientation within the bilayer is much slower, typically of the order of tens of milliseconds. Furthermore, a bilayer has two distinct diffusion modes, namely, the lateral diffusion within the bilayer plane, and the motion of lipid molecules from one monolayer leaflet to the other, the so-called flip-flop process. The latter diffusional mode is extremely slow, of the orders of hours, possibly days,

The fast lateral diffusion of lipids in the plane of the membrane is a typical liquid property. For a typical size of a cell, a lipid molecule can travel along the whole cell membrane within less than half a minute. Lipid lateral diffusion depends, of course, on the temperature as well as on the state of matter of the lipid bilayer. If the lipid membrane is in the solid state (gel phase, see Figures 6.7 and 6.11) all dynamical processes slow down significantly, and lateral diffusion is slowed down at least a hundred times.

The diffusion in lipid bilayers can be monitored by a number of experimental techniques, such as fluorescence correlation spectroscopy and pulsed field gradient NMR methods (pfg-NMR). The latter method has the advantage of being non-invasive and is the only one where no probe molecules are necessary. The lipid lateral diffusion coefficient (D_I) is directly obtained from a sample of stacked lipid bilayers. The method can also be used to study lipid domains in bilayers (see Section 6.3.6). Table 6.3 presents lateral diffusion coefficients obtained by pfg-NMR for different systems. Here, it can also be seen that the lateral diffusion depends on the packing of the lipids in the bilayer; the tighter the packing (smaller cross-sectional head group area), the smaller the diffusion coefficient. The rate of lateral diffusion increases with increasing number of double bonds, as a consequence of the increased head-group area caused by the unsaturation. It can also be inferred from Table 6.3 that D_L decreases in the order DOPC>POPC>DPPC>DMPC, again in line with the decrease in the head-group area, mainly caused by the degree of saturation/unsaturation of the acyl chains. Moreover, egg sphingomyelin (eSM) has a lower D_L than DPPC, and dioleoylphosphatidylglycerol (DOPG) with its repulsive charged head-group exhibits a larger diffusion coefficient than DOPC, once more in agreement with what can be expected from the head-group area in these bilayer systems.

Finally, the effect on D_L upon a change in lipid packing is shown by the classical systems of a mixture of phospholipids and cholesterol, and the "condensing effect of cholesterol". In Table 6.3, this is illustrated by an increasing content of CHOL in DMPC bilayers. The flat, rigid ring system of cholesterol is effective in tightening the packing of the hydrocarbon chains, resulting in a reduced lipid lateral diffusion.

This will be discussed in some detail in Section 6.3.6, where so-called lipid rafts are studied using NMR methods.

6.3 Lipids Play a Fundamental Role in Membrane Function

Biological membranes contain many different lipids and many of them do not form a bilayer on their own. Why do organisms use energy to synthesize lipids that do not form

TABLE 6.3 Headgroup Areas from SAXS Experiments and Corresponding D_L for Somelipids in Bilayers

| | Fatty Acid Chain | | | |
|----------------------------|------------------|-----------------------|--------------------------|-----|
| Lipid | Composition | Area, nm ² | D_L , $\mu m^2 s^{-1}$ | t°C |
| DOPC | 18:1/18:1 | 0.72 | 8.25 | 25 |
| POPC | 16:0/18:1 | 0.68 | 7.79 | 25 |
| DPPC | 16:0/16:0 | 0.64 | 17.8 | 45 |
| DMPC | 14:0/14:0 | 0.61 | 5.82 | 25 |
| SOPC | 18:0/18:1 | 0.63 | 6.6 | 25 |
| SLPC | 18:0/18:2 | 0.66 | 8.2 | 25 |
| SAPC | 18:0/20:4 | 0.68 | 10.0 | 25 |
| SDPC | 18:0/22:6 | 0.70 | 11.2 | 25 |
| eSM | | 0.53 | 4.5 | 50 |
| DOPG | 18:1/18:1 | 0.80 | 15 | 30 |
| A. laidlawii lipid extract | | ca 0.6 | 2.7 | 30 |
| CHOL/(CHOL+DMPC) | | | | |
| 0.0 | | 0.61 | 9.0 | 30 |
| 0.1 | | 0.53 | 4.9 | 30 |
| 0.2 | | 0.48 | 2.6 | 30 |
| 0.3 | | 0.44 | 2.0 | 30 |

The fatty acid chain composition is given as C:D, where C and D are the number of carbon atoms and double bonds in a single chain, respectively. Two values are given for each lipid — one for each of the fatty acid tails. Note that for DPPC the diffusion coefficent is measured at 45°C and for eSM at 50°C, for these systems to be in the fluid bilayer state ($T_{\rm m}$ is around 41°C).

DOPC is dioleoylphosphatidylcholine; POPC is palmitoyloleoylphosphatidylcholine; DPPC is dipalmitoylphosphatidylcholine; DMPC is dimyristoylphosphatidylcholine; SOPC is stearoyloleoylphosphatidylcholine; SLPC is stearoyllinoleoylphosphatidylcholine; SAPC is stearoylarachidonoylphosphatidylcholine; SDPC is stearoyldocosahexaenoylphosphatidylcholine; eSM is egg sphingomyelin; DOPG is dioleoylphosphatidylglycerol; CHOL is cholesterol. (Data from Lindblom G, Orädd G (2009) Lipid lateral diffusion and membrane heterogeneity. Biochim Biophys Acta 1788: 234-244.)

bilayers, necessary for a functioning and tight, non-leaky cell membrane? It is essential for the membrane to establish a proper barrier to the surroundings. Evidently, there must be other reasons for having appreciable amounts of lipids that, when separated and dispersed in water, form other aggregate structures than bilayers. Important advances in the understanding of the lipid composition of biological membranes have been made during the last 20 to 30 years.

We have seen above that the aggregate structure a lipid forms depends on variables like temperature, composition, molecular structure, etc. Below, we will discuss the physico-chemical properties of lipids, how organisms adjust the lipid composition in their cell membranes to changes in the environmental conditions of the cells, and what determines the lateral organization of the membrane lipids, i.e. domain formation and the importance of cholesterol.

6.3.1 Regulation of Membrane Lipid Composition

It is well documented that all kinds of organisms adapt their membrane lipid composition to the prevailing environmental and physiological conditions. This is necessary to keep a stable, non-leaky lipid bilayer. Cells seem to use any of three strategies to change the geometry or physico-chemical properties of the lipid molecules in the bilayer: (i) changes in the acyl chain structure; (ii) changes in the polar head-group structure; and (iii) reshuffling of acyl chains to form new lipid species without changing the average acyl chain composition.

6.3.1.1 Escherichia coli

The gram-negative bacterium *E. coli* is recognized as one of the foremost prokaryotic model organisms. It has only three main membrane phospholipids that occur frequently in prokaryotic as well as eukaryotic organisms. The regulation of the lipid composition in wild-type cells is brought about by changes in the acyl chain structure, above all in the degree of unsaturation of the acyl chains. This is a very common response among a variety of organisms to changes in the environmental temperature. In this way *E. coli* succeeds in maintaining the lipids in a bilayer state and avoids the formation of either crystalline or non-lamellar liquid crystalline phases.

Phosphatidylethanolamine (PE), phosphatidylglycerol (PG) and diphosphatidylglycerol (DPG) are the main membrane lipids synthesized by wild-type $E.\ coli$. PE has the strongest propensity to form reversed non-lamellar phases and this ability is profoundly influenced by the length and degree of unsaturation of the acyl chains. Wild-type $E.\ coli$ cells synthesize all their fatty acids themselves and these are not incorporated from the growth medium. When the growth temperature of $E.\ coli$ is increased, the polar head-group composition remains practically constant, while the saturation of the acyl chains is increased (Figure 6.22). Generally, an increased temperature shifts the membrane lipid phase equilibria from lamellar toward cubic and/or hexagonal (H_{II}) phases. Wild-type $E.\ coli$ cells respond to higher growth temperatures by incorporating shorter and more saturated acyl chains into their membrane lipids. Such changes decrease the ability of PE to form a non-lamellar phase, thus counteracting the increase in temperature (Figure 6.22).

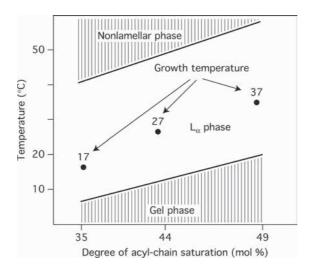


Fig. 6.22 ■ The bacterium *E. coli* grows in a "window" of temperature and acyl-chain saturation. It has to change the fatty acyl chain saturation to "solve" the problem of keeping a balance between non-lamellar- and lamellar-forming lipids. (Adapted with permission from Lindblom G, Orädd G, Rilfors L, Morein S. (2002) Regulation of lipid composition in Acholeplasma laidlawii and Escherichia coli membranes: NMR studies of lipid lateral diffusion at different growth temperatures. Biochemistry 41: 11512–11515. Copyright (2002) American Chemical Society.)

6.3.1.2 Acholeplasma laidlawii

The regulation of the membrane lipid composition by the cell wall-less bacterium Acholeplasma laidlawii strain A has been intensively studied. This organism can be grown under conditions where the regulatory changes occur predominantly in the polar headgroup structures. In this way, the cells strive to maintain a certain balance between the lipids constituting a bilayer and those forming a reversed non-lamellar structure. Therefore, it provides an excellent model for physico-chemical investigations of an intact cell membrane mainly for two reasons: (i) its ability to introduce controlled changes into the membrane acyl chains and introduce sterols and other molecules into the membrane and (ii) the ease with which pure membranes free from contaminants can be obtained. Early differential scanning calorimetry studies showed that the lipids in the membrane of A. laidlawii exhibited a reversible gel-to-liquid crystalline phase transition. Furthermore, from X-ray scattering and ²H NMR spectroscopy it was concluded that a fluid, lipid bilayer structure built these membranes. The metabolic regulation and the phase equilibria of the membrane lipids follow a different route than for E. coli, and yet they produce the same consequences concerning the physico-chemical properties of the membrane.

The dominant membrane lipids in mycoplasmas like A. laidlawii are the glucolipids, monoglucosyldiacylglycerol (MGlcDAG) and diglucosyldiacylglycerol (DGlcDAG), lesser amounts of phosphoglucolipids with phosphate groups esterified to the sugar

Fig. 6.23 ■ Structure of the gluco- and phospho-glucolipids in the A. laidlawii membrane: 1.MGlcDAG;2.MAMGlcDAG;3.DGlcDAG;4.MADGlcDAG;5.GPDGlcDAG;6.MABGPDGlcDAG. (Reprinted with permission from Andersson A-S, Rilfors L, Bergqvist M, Persson S, Lindblom, G. (1996) New aspects on membrane lipid regulation in Acholeplasma laidlawii A and phase equilibria of monoacyldiglucosyldiacylglycerol. Biochemistry 35: 11119–11130. Copyright (1996) American Chemical Society.)

head groups (Figure 6.23), and finally also phosphatidylglycerol (PG) (see Figure 6.5). Thus, A. laidlawii strain A synthesizes a total of seven membrane lipids. Three of the lipids are able to form non-lamellar phases: MGlcDAG, monoacyl-MGlcDAG (MAMGlcDAG), and monoacyldiglucosyldiacylglycerol (MADGlcDAG). The lamellarforming lipids are PG, DGlcDAG, glycerophosphoryl-DGlcDAG (GPDGlcDAG), and monoacylbisglycerophosphoryl-DGlcDAG (MABGPDGlcDAG). GPDGlcDAG may also form normal micelles at high water contents.

A. laidlawii can be grown under conditions where fatty acids cannot be endogenously synthesized, and the cells are, therefore, forced to incorporate the exogenously supplied fatty acids into their membrane lipids. The cells respond by adjusting the composition of the polar head-groups to the incorporated acyl chains, and the polar head-group

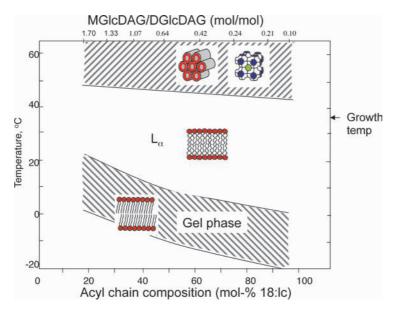


Fig. 6.24 ■ The bacterium A. laidlawii grows when its lipids are in a lamellar state. There is a balance between the non-lamellar and lamellar-forming lipids in the "window" shown in the figure. Note also that the upper phase transition to a non-lamellar state is almost constant in temperature. The bacterium grows at a temperature that is about 10–15° below this transition. The packing of the lipids in the membrane seems to be of crucial importance. (Adapted from Lindblom G, Rilfors L. (1989) Cubic phases and isotropic structures formed by membrane lipids — Possible biological relevance. Biochim Biophys Acta 988: 221–256. Copyright (1989) Elsevier.)

composition is regulated in a coherent way. Generally, the fraction of the lipids forming reversed non-lamellar structures decreases when the length and the unsaturation of the acyl chains are increased. The regulation of the ratio between the lipids forming lamellar and non-lamellar phases yields phase transition temperatures from a lamellar to a nonlamellar phase within a rather narrow interval for total lipid extracts (Figure 6.24). Note also that the unsaturation and length of the acyl chains play a critical role for the survival of the bacterium (Figure 6.25). With short, saturated acyl chains (lower left corner in Figure 6.25) even a new lipid, only observed in this harsh environment, has to be synthesized in order to keep the critical balance between lamellar and non-lamellar-forming lipids, indicating that the packing of the lipids in the membrane is fundamental to its functioning.

An important conclusion from the studies of A. laidlawii and E. coli is that the cells always seem to adjust the membrane lipid composition so that a lamellar liquid crystalline phase is maintained, thus avoiding the formation of either a gel or non-lamellar liquid crystalline phase (Figures 6.24 and 6.22). The relationship between the membrane lipid composition and the physico-chemical properties of the lipids has also been explored

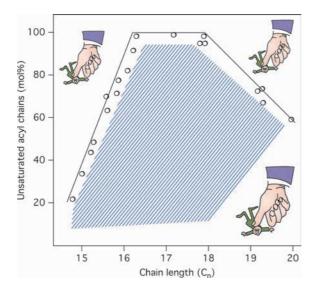


Fig. 6.25 ■ Acyl chain unsaturation as a function of the chain length in the membrane of surviving A. laidlawii bacteria. A. laidlawii will survive only within the blue-hatched area, where the lipid chain length and unsaturation pertain to an optimal packing of the lipids and the balance between lamellar- and non-lamellar-forming lipids gives the correct physico-chemical properties in the plasma membrane. On the left hand side of the blue-hatched area, i.e. where the acyl chain is very short, the bacterium has to synthesize a new lipid with three acyl chains, MADGlcDAG (Figure 6.23) to survive. In a region with a slightly longer acyl chain, between C_{16} and C_{17} , and at a low unsaturation, another lipid is synthesized, namely, MAMGlcDAG. (Data from Andersson A-S, Rilfors L, Bergqvist M, et al. (1996) New aspects on membrane lipid regulation in Acholeplasma laidlawii A and phase equilibria of monoacyldiglucosyldiacylglycerol. Biochemistry 35: 11119–11130, and Wieslander Å, Nordström S, Dahlqvist A, et al. (1995) Membrane lipid composition and cell size of Acholeplasma laidlawii strain A are strongly influenced by lipid acyl chain length. Eur J Biochem 227: 734-744.)

in other prokaryotic organisms. Like A. laidlawii and E. coli, Clostridium butyricum and Bacillus megaterium seem to regulate their membrane lipid composition.

It seems implicit that the temperature acclimation or adaptation modifies the phase transition temperatures of the lipids so that a fluid bilayer is always stable and the growth temperature is confined to a "window" bounded by the gel and non-lamellar phases.

6.3.1.3 Eukaryotic organisms

Eukaryotic organisms, like fungi and poikilothermic animals (animals whose body temperature follows the environment) often change their membrane lipid composition in response to changes in the ambient temperature. A similar regulation of the membrane lipid composition occurs in plants. The freezing intolerance of the plasma membrane of oat and rye leaves is primarily a consequence of membrane destabilization resulting from freeze-induced dehydration. The lipid composition of the plasma membrane isolated from leaves of spring oat is vastly different from that of winter rye. The plasma membrane of spring oat contains large fractions of phospholipids, cerebrosides and acylated sterylglucosides, while that of winter rye has a greater proportion of phospholipids, and a lesser fraction of cerebrosides, but larger fractions of free sterols. However, for both organisms, cold acclimation results in an increase of the fraction of phospholipids, and a decrease in the cerebroside component. It turns out that the lipids form an H_{II} phase in the freezing injury. The incidence of the H_{II} phase correlates with lethal injury to both protoplast and leaf tissue as indicated by loss of osmotic responsiveness of protoplasts and leakage of the intracellular contents of leaves. The temperature dependence for the onset of the freeze-induced formation of the H_{II} phase is significantly different for winter rye and spring oat and it is, not surprisingly, associated with the differences in the lipid composition of the plasma membranes. However, freeze-induced formation of the $H_{\rm II}$ phase does not occur after cold acclimatization in either of the organisms because of the strong decrease in the cerebroside fraction in the plasma membrane. Such an adaptation is understandable given that cerebrosides promote the formation of the H_{II} phase at low temperatures. The polar head group of cerebrosides has a low level of hydration, resulting in a more wedge-shaped molecule that easily packs into an H_{II} structure. Therefore, the propensities of the plasma membranes of rye and oat to undergo the lamellar to H_{II} phase transition during freeze-induced dehydration appears to be a consequence of the physico-chemical properties of the membrane lipids, including bilayer surface hydration and lipid packing.

6.3.2 Role of Non-lamellar-forming Lipids for Membrane Function

The lamellar liquid crystalline phase with its multi-bilayer structure has long been used as a model for biological membranes, and single-bilayer vesicles are frequently used in pharmaceutical applications. The awareness of the formation of non-lamellar structures has gradually changed the view on the functional role played by membrane lipids in cellular processes. There is now a great deal of experimental and theoretical evidence showing that they actively participate in many important functions of the cell. Surprisingly little has been written on this aspect of lipids in biochemistry textbooks so far.

6.3.2.1 Special membrane structures

One reason for cells to synthesize non-lamellar-forming membrane lipids, and to maintain a given balance between these lipids and the lamellar-forming ones, is that nonlamellar-forming lipids are needed to form either non-bilayer structures, or bilayer structures with a small radius of curvature. Non-bilayer structures have been implicated in the fusion and fission of lipid bilayers. Bilayer structures with a small radius of curvature occur in several types of biological membranes, such as the endoplasmic reticulum, the inner mitochondrial membrane, and the grana stacks of thylakoid membranes in chloroplasts (Figure 6.2). In particular, it has been shown by transmission electron micrographs of the smooth endoplasmic reticulum and the inner mitochondrial membrane that these resemble bicontinuous cubic or L_3 phase structures (see Figures 6.16 and 6.17). A highly ordered, branched tubular membrane structure called the prolamellar body is present in etioplasts, organelles found in leaves of plants grown in the dark. After exposure to light, the etioplasts transform into chloroplasts and the prolamellar body develops into the thylakoid membranes of the chloroplast.

6.3.2.2 Influence on the activity of membrane-associated proteins

Investigations from the 1980s showed that the efficiency of protein incorporation during reconstitution into vesicles and the activity of membrane proteins are both enhanced in the presence of lipids forming non-bilayer structures, or by the incorporation into the membrane of molecules known to destabilize the bilayer structure. The activity of the phosphatidylcholine-specific phospholipase C from *Bacillus cereus* is enhanced by the presence of lipids that destabilize the lamellar phase. This activation is attributed to a packing stress in the lipid bilayer ("frustration"), rather than to the actual formation of reversed non-lamellar phases. Such "frustrated" lipid bilayers can play a role in the anchorage and activation of peripheral membrane proteins (Section 4.4.1). For protein kinase C, both the partitioning of the enzyme to a membrane and the activity of the membrane-bound form of the enzyme are increased in the presence of non-lamellar-forming lipids. The fungal peptide alamethicin forms a voltage-dependent ion channel in membranes, and the states of higher conductance are more probable when the fraction of a non-lamellar-forming lipid in the bilayer is increased i.e. when the spontaneous curvature of the two monolayers is increased.

6.3.3 Membrane Fusion and Fission

Membrane fusion is a very important phenomenon in all cells. It occurs when two initially separate and apposed membranes merge into one by undergoing a sequence of intermediate transformations. It is involved in membrane trafficking or vesicle-mediated transport, sperm-egg fusion and virus-cell fusion. In particular, the trafficking of proteins in the secretory pathway inside the cell is mediated by vesicles and relies on various fusion and fission processes, such as transport from the endoplasmic reticulum (ER) to the Golgi apparatus. The molecular mechanisms behind the processes of fusion and

fission have been studied extensively both experimentally and theoretically over the last decades. Although the details are still not fully recognized, the hypothesis is that all fusion is essentially lipidic at its core. Here we will confine ourselves to processes occurring in simple lipid systems, in particular, systems where non-lamellar structures are suggested to be involved in membrane fusion and fission. The essential step in fusion is a rearrangement of the lipid molecules from two apposed membranes to form a single, continuous membrane. For an understanding of the fusion process, both dynamical and structural aspects have to be considered.

Early on it was proposed that a reorganization of the membrane structure upon membrane fusion required that the lipid bilayer be broken up to form other aggregate structures. Lysophosphatidylcholine (LPC) can induce fusion between erythrocytes, as well as between erythrocytes and fibroblasts, and LPCs are known to form normal micellar solutions at high water contents due to their low packing parameter (cf. Figure 6.13). Non-lamellar aggregate structures have been seen in the fusion of intact erythrocytes upon addition of appropriate agents. Structures related to the transitions between lamellar and $H_{\rm II}$ or cubic phases are involved in membrane fusion, and it has been shown that low levels of particular lipids like lysolipids, or some types of fusion peptides or proteins, can significantly enhance the rate of fusion in model membranes and biomembranes (Figure 6.26).

Lipid membrane fusion starts with a close contact between the outer lipid monolayers of the two fusing membranes, while the distal monolayers remain separate. The initial lipid bridge between the membranes is referred to as the fusion stalk (Figure 6.26) and signifies the first stage of fusion called hemifusion. The membrane stalk is a neck-like structure in which only the outer monolayers of the two fusing bilayers are connected. During the stalk formation, regions of high membrane curvature are formed, possibly mediated by local non-lamellar structures. At some stage, the two inner monolayers make contact and an aqueous pore is formed connecting the interior regions of the two vesicles.

Lipid membrane fusion is observed only for specific lipid compositions and specific ions in the aqueous bathing solution or upon dehydration of the intramembrane contact. In the initial state, the membrane monolayers accumulate energy, which is released upon fusion. The fusion-driving energy will increase if the curvature of the contacting membrane monolayers differs from their spontaneous curvature (see Section 6.3.1).

Membrane fission, i.e. division of an initially continuous membrane into two separate ones, proceeds via the formation of a membrane neck, which is reminiscent of a fusion pore. Thus, fission begins with self-merging of the inner monolayer of the neck membrane, which generates a fission stalk analogous to the fusion stalk. Subsequent selfmerging of the outer monolayer of the lipid membrane neck completes the fission process.

Proteins can affect the lipid composition of the contacting monolayers to produce negative spontaneous curvature. Hence, phospholipases and acyltransferases that initiate enzymatic cascades leading to increased concentrations of such lipids as diacylglycerol (DAG) and PE may promote fusion and have, indeed, been implicated in some intracellular fusion reactions. Fusion proteins can cause distortions of the bilayer

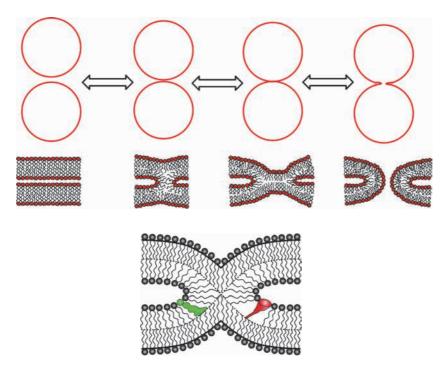


Fig. 6.26 ■ *Top*: Steps in the fusion and fission of membranes. The third stage is where the stalk is formed. The shapes of the lipids are of great importance for formation of the stalk (see the section about bilayer curvature, Section 6.2.1). Bottom: A stalk structure proposed to occur in membrane fusion. The dashed lines show the bilayer midplane. A green-colored cone-shaped lipid molecule such as PE (on the left wing of the stalk) matches the stalk structure, whereas the red-colored inverted-cone-shaped lipid molecule such as LPC (on the right wing of the stalk) disturbs the packing of the lipids in the stalk.

structure, which leads to the elastic stresses that cause fusion. For example, fusion mediated by influenza hemagglutinin (HA) critically depends on a specific boomerang-like conformation of the membrane-inserted fusion peptide domains that is hypothesized to produce a bilayer distortion required for the hemifusion.

There are a number of proteins that are central for fusion of cell membrane vesicles, called SNARE proteins (soluble NSF (N-ethylmaleimide-sensitive factor)-attachment protein receptors). The primary role of SNARE proteins is to mediate vesicle fusion through exocytosis. SNAREs can be divided into two categories: vesicle or v-SNAREs, which are incorporated into the membranes of transport vesicles during budding, and target or t-SNAREs, which are located in the membranes of target compartments. The v-SNARE and the t-SNARE interact and get twisted forming a so called "leucine zipper" (see Section 3.3.2) that pulls the merging membrane together with eventually fusion.

Finally, it should be clear that the mechanisms behind fusion of cell membranes and vesicles are not yet fully understood on the molecular level, and several different ideas

have been proposed on this difficult matter. However, it is clear that certain fusion peptides and proteins are involved, e.g. SNARE proteins, proteolipid complexes and receptors activated by calcium ions. Furthermore, fusion requires a local rearrangement of the lipids in the involved membranes to allow for regions of very high curvature. Especially, the formation of reversed liquid crystalline intermediate structures may occur for topological reasons. Therefore, lipids with a large propensity for forming H_{II} or reversed cubic structures, such as PE lipids, will facilitate fusion processes whereas lamellar-forming PC lipids will not.

6.3.4 Lipid Synthesizing Enzymes

The structure of fatty acid synthase is described in Section 8.4. In mammalian cells, the endoplasmic reticulum (ER), where the different synthesizing enzymes reside, is the main lipid factory, in which the bulk of phospholipids and sterols, as well as substantial amounts of storage lipids such as triacylglycerol and steryl esters, are produced. In addition, the ER synthesizes ceramide, the precursor of all sphingolipids. Furthermore, the ER supplies a large portion of membrane lipids to the Golgi and plasma membrane because these distal secretory organelles have little or no capacity to produce their own. Despite an extensive exchange of material by membrane trafficking (for instance, by lipid vesicles), the ER and plasma membrane show remarkable differences in their lipid composition, and sterols are rare in the ER but abundant in the Golgi and plasma membrane.

The regulation of the membrane lipid composition implies that the activity of the enzymes synthesizing the lipids (lipid synthases) is adjusted to the prevailing growth conditions of the cells. Some kind of signal(s), reflecting the status of the lipid bilayer, must thus be transferred from the bilayer to the lipid synthases. The lipid synthases are generally tightly associated to the lipid bilayer, and one possibility is that the activity of these enzymes is directly influenced by the properties of the lipid bilayer (see Figures 6.19 and 6.27). Another alternative is that effector molecules binding to the synthase regulate its activity. These effector molecules can consist of membrane lipids, or of a special protein that in turn senses the status of the lipid bilayer.

6.3.4.1 CTP:phosphocholine

PC is a major membrane lipid in most eukaryotic cells. Therefore, the regulation of the activity of the enzyme CTP:phosphocholine cytidylyltransferase (CCT) is very important for membrane biogenesis. The activity of CCT is increased by anionic phospholipids and by neutral lipids like diacylglycerol. An amphipathic α-helical peptide (a three-fold repeat of 11 amino acid residues) of CCT binds to the surface of anionic lipid vesicles, and the activating effect of anionic lipids is attributed to an electrostatic interaction between these lipids and basic amino acid residues in the amphipathic helix.

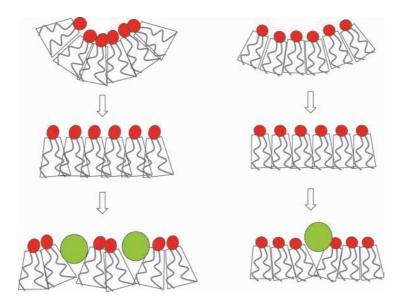


Fig. 6.27 • A cartoon showing how the cell may regulate membrane elastic stress upon binding of the amphipathic helix of CCT (green; viewed down the helix axis) to a DOPE (left) in comparison to a DOPC (right) bilayer. More CCT molecules are needed to relax (flatten) a DOPE bilayer than a DOPC bilayer, since the molecular shape for the two lipid molecules differs (cf. Figure 6.13). (Adapted with permission from Attard GS, Templer SH, Smith WS, et al. (2000) Modulation of CTP: phosphocholine cytidyltransferase by membrane curvature elastic stress. Proc Natl Acad Sci USA 97: 9032–9036. Copyright (2000) National Academy of Sciences, USA.)

The activity of CCT is modulated by the stored curvature or elastic stress in the monolayers of a lipid membrane. It increases monotonically when the elastic stress in the lipid monolayers increases (Figure 6.27). In contrast, the enzyme activity decreases significantly by incorporation of small fractions of detergent molecules (cf. Figure 6.12) into the bilayer, i.e. when the spontaneous curvature of the monolayers is decreased. Thus, a purely physical feedback signal can play a key role in the regulation of membrane lipid synthesis. However, the molecular details remain unknown.

The model suggests that the stored elastic energy of the lipid bilayer modifies the activity of curvature-sensitive enzymes through the interaction with amphipathic α -helices. As their binding depends on the lipid composition, this results in a biophysical feedback mechanism for the regulation of the stored elastic energy that depends on the packing of the lipids in the bilayer. Thus, restrictions are imposed on the balance between lamellar- and non-lamellar-forming lipids in the plasma membrane and on the concentrations of particular lipids. By using measured values of lipid curvatures from A. laidlawii the theoretical model gives quite a good, although as yet not fully quantitative, description of the membrane process (see Further Reading).

6.3.5 BAR Domains Order Membrane Curvature

The interior of a eukaryotic cell contains a multitude of structures such as vesicles, tubules and disks, in which the membranes are folded in defined and dynamic geometries. Understanding how membrane shape is regulated within the cell is a fundamental problem of contemporary cell biology.

Peripheral membrane proteins are emerging as important players in membraneremodeling phenomena. Such proteins possess different classes of membrane-binding domains, like the so-called BAR domains (Bin/amphiphysin/Rvs; amphiphysin is a brain-enriched protein with an N-terminal lipid interaction). These are highly conserved protein dimerization domains. The BAR domain is banana-shaped (Figure 6.28) and binds to membranes via its concave face. It is capable of sensing membrane curvature by binding preferentially to curved membranes. The domain is found in a large family of proteins that are able to tubulate lipid membranes.

Many BAR proteins contain alternative lipid specificity domains that help target these proteins to particular membrane compartments. Some also have SH3 domains that bind to dynamin (and thus proteins-like amphiphysin and endophilin — a related protein, see Figure 6.28), so are implicated in the orchestration of vesicle scission. Therefore, these proteins are involved in the processes of endocytosis that implies massive sculpting and trafficking of membranes.

BAR domains induce membrane curvature in an organelle-specific manner in live cells on a timescale of seconds. This curvature is dependent on the unique characteristics of each BAR subtype. These differential effects are intriguing in terms of how precisely this bending occurs and how the resulting morphology relates to endogenous cellular events. A number of different mechanisms are responsible for the resulting morphology of cells and organelles. Lipid packing, lipid composition, integral membrane protein localization or wedge-like insertion of peripheral membrane proteins, protein crowding, protein scaffolding and cytoskeletal-based mechanisms contribute to curvature of intracellular membranes.

Thus, BAR domain superfamily proteins have emerged as central regulators of dynamic membrane remodeling, thereby playing important roles in a wide variety of cellular processes, such as organelle biogenesis, cell division, cell migration, secretion and endocytosis.

6.3.6 Lipid Domains and Rafts in Membranes

In 1972, Singer and Nicolson launched their classical model of the membrane as a matrix in which the proteins have a degree of motional freedom in a lipid "sea". This "fluid mosaic model" became the framework and benchmark for our current understanding of membrane bilayers and their physiological function (Figure 6.29).

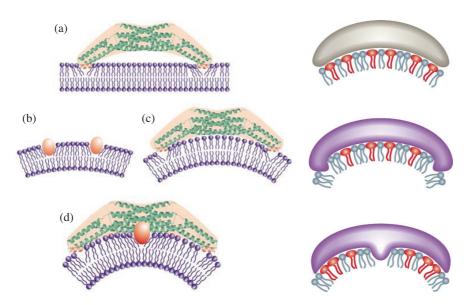


Fig. 6.28 ■ Left: Two potential mechanisms for driving membrane curvature by BAR proteins. (a) Kissing adhesion of a BAR domain on planar lipid bilayer. The amphiphilic helices at the ends of the banana-like protein are essential for the membrane binding, leading to insertion of the helices into the lipid bilayer interface. (b) Insertion of hydrophobic portions of macromolecules into one leaflet can create bilayer surface discrepancy that causes membrane curvature. (c) The simple N-BAR domain, such as amphiphysin, induces membrane curvature by impressing the concave surface onto the membrane. The rigidity of the molecule is required for this mechanism. (d) To drive membrane curvature, the endophilin BAR domain uses both the rigid crescent-shape-mediated deformation and the insertion of hydrophobic ridge on the concave surface in addition to kissing adhesion of N-BAR to membrane surface. (Reprinted with permission from Masuda M, Takeda S, Sone M, et al. (2006) Endophilin BAR domain drives membrane curvature by two newly identified structure-based mechanisms. EMBO J 25: 2889-2897. Copyright (2006) European Molecular Biology Organization. Right: Mechanisms of membrane shaping by BAR domains. Top: Scaffolding mechanism promoted by binding to inverted cone-shaped negatively charged lipids (red). Middle and bottom: Hydrophobic insertion mechanisms embedding protein parts into the lipid layer. (Reprinted with permission from Qualmann B, Koch D, Kessels MM. (2011) Let's go bananas: Revisiting the endocytic BAR code. EMBO J 30: 3501-3515. Copyright (2011) European Molecular Biology Organization.

However, the homogeneous nature of the membrane proposed in this model, characterized by the random distribution of molecular components in the membrane, was later altered. Many recent studies have revealed that cell membranes possess a rather complex lateral organization. For example, it was discovered by single-particle tracking techniques that labeled lipid or protein molecules perform a lateral diffusive motion, but that they can get temporarily confined into discrete domains on the membrane.

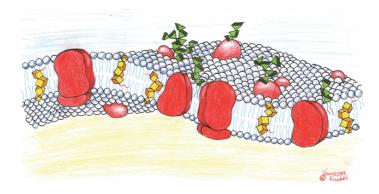


Fig. 6.29 ■ A cartoon of the fluid mosaic model of a biological membrane from 1972 according to Singer and Nicolson. The yellow transmembrane molecules represent cholesterol and the green parts sticking out into solution represent sugar molecules. Membrane proteins are colored red. (Courtesy of Vanessa Kunkel.)

Lateral domains in membranes enriched in cholesterol, sphingolipids and phospholipids (referred to as lipid rafts) are a central area of research in lipid biology (Figure 6.30). These nanosized raft domains, occurring in fluid membranes, are suggested to play a dominant role in signal transduction, lipid trafficking, regulation of the activity of membrane proteins and transcytosis (the vesicular transport of macromolecules from one side of a cell to the other).

Lipid domain formation was first observed in detergent resistant membranes (DRM) that could withstand solubilization by the detergent Triton X-100. DRMs were enriched in cholesterol and the mainly saturated lipid sphingomyelin, while the remainder of the membrane contained more unsaturated lipids, mainly phosphatidylcholines. The evidence for the existence of raft structures in different cell membranes is steadily increasing, although most of it is indirect. Thus, the existence of stable rafts in biological membranes is still under intense scrutiny and debate. One of the problems is that the membrane rafts are probably too small to be resolved by the most common techniques used here, like conventional fluorescence microscopy, since their lengthscales are below the diffraction limit of optical microscopy. Examples from native lung membranes are shown in Figure 6.31.

However, in recent years new microscopy methods have been developed, called superresolution fluorescence microscopy, by which very small domains down to 20-200 nm or smaller can be detected. Although attempts to observe rafts on living membranes have been performed with these new fluorescence techniques, it seems that a cell membrane might be a too complex system and we are still awaiting a clear-cut picture of a membrane raft (see Further reading). Thus, it is difficult to monitor physical properties of the rafts, such as their size, lifetime, dynamics and lipid composition

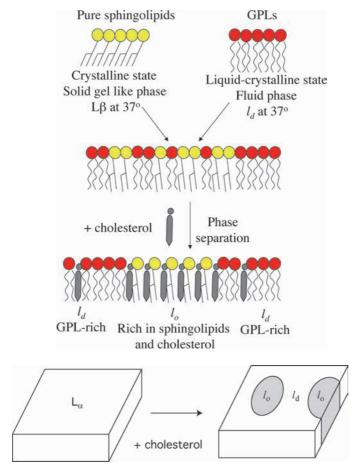


Fig. 6.30 ■ *Top*: Cholesterol induces lateral phase separation into domains in the membrane containing glycerophospho- and sphingo-lipids. GPL stands for glycerophospholipid. The grey-yellow part of the membrane at the bottom represents a raft. *Bottom*: An illustration of the formation of domains or rafts in a fluid lipid bilayer upon the addition of cholesterol. Note the smooth circular cross-sections that occur, since the domains, l_o , in the fluid bilayer are also fluid. L_α stands for a lamellar phase, l_o is the ordered lipid domain and l_d is the disordered phase.

directly on living cells. Investigations on lipid model membranes, on the other hand, are very informative and the existence of lipid domains or lateral phase separation is widely accepted.

Lipid domains have been observed in multi-bilayer lipid systems, lipid monolayers or lipid (giant) vesicles by atomic force microscopy and fluorescence microscopy, fluorescence quenching, single-particle tracking, differential scanning calorimetry, solid state NMR and NMR diffusion methods, and X-ray diffraction. Generic phase diagrams constructed from such investigations are shown in Figure 6.32. A detailed phase diagram is of great importance to understand the driving forces behind domain formation

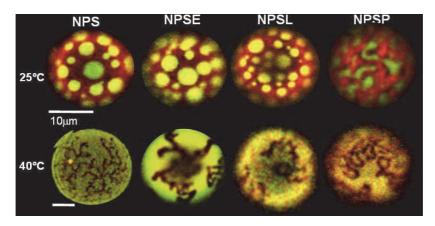


Fig. 6.31 ■ Effect of temperature on phase segregation in native pulmonary surfactant membranes and membranes reconstituted from some of its fractions. Segregation of phases was observed at temperatures below (25°C) and above (40°C) the thermotropic transition of surfactant membranes, by confocal microscopy of GUVs made from native pulmonary surfactant (NPS) or its fractions (NPSE, lipid fraction containing all lipid species and hydrophobic proteins, NPSL, lipid fraction without proteins, and NPSP, lipid fraction without proteins and cholesterol), loaded with 0.1 mol% each of the probes DiIC18 (red) and Bodipy-PC (yellow). The scale bar represents 10 µm for all images except that of the GUV made from NPS and observed at 40°C, which has its own smaller scale bar. (Reprinted with permission from Bernardino de la Serna J, Orädd G, Bagatolli LA, et al. (2009) Segregated phases in pulmonary surfactant membranes do not show coexistence of lipid populations with differentiated dynamic properties. Biophys J 97: 1381–1389. Copyright (2009) Biophysical Society.)

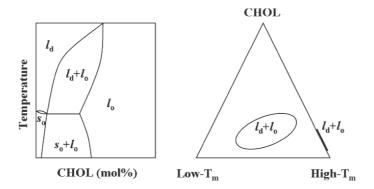


Fig. 6.32 ■ Generic temperature/concentration phase diagrams of a binary system with a saturated phospholipid, cholesterol (CHOL) and water (left), and a ternary system with saturated lipid (Low- $T_{\rm m}$), unsaturated lipid (High- $T_{\rm m}$), CHOL, and water (right). In the ternary phase diagram only the fluid-fluid phase coexistence areas are depicted. Note the two-phase area containing disordered l_d and ordered l_o lamellar phases. s_o stands for the gel phase. $T_{\rm m}$ is the temperature at the phase transition from a gel to a lamellar phase. (Adapted with permission from Lindblom G, Orädd G. (2009) Lipid lateral diffusion and membrane heterogenity. Biochim Biophys Acta 1788: 234-244. Copyright (2009) Elsevier.)

in lipid bilayers. In fact, it seems that most relevant lipid systems to date show a similar phase behavior to the one first published for the system with dipalmitoylphosphatidylcholine (DPPC), cholesterol (CHOL) and water. In this phase diagram, it was discovered that liquid-ordered (l_o) and liquid-disordered (l_d) lamellar phases formed in equilibrium with each other. Both these phases are in a fluid liquid crystalline state, but the hydrocarbon chains in the l_o phase are more ordered, or stretched than those in the l_d phase. The phase diagram exhibits a large two-phase area with the l_o and the l_d phases. See also the previous discussion of ternary phase diagrams (Figure 6.8b). The rafts in cell membranes are believed to be made up of such an l_o phase structure (see Further Reading).

Rafts are believed to contain high levels of cholesterol and sphingolipids as well as saturated phospholipids. The presence of sphingomyelin or glycosphingolipids together with cholesterol promotes ordering of the lipid hydrocarbon chains, and this led to the suggestion that rafts have a structure similar to the l_o phase. Let us briefly examine how one can detect lipid domains in bilayers by the NMR diffusion method (pfg-NMR), and from the data get part of a phase diagram. The influence of domains on the lipid lateral diffusion enables the lateral phase separation in lipid bilayers to be determined by NMR. Lipids will diffuse either into or out from the separated phases in the fluid bilayer by some exchange mechanism or, provided that the border between different domains presents an obstacle to lipid diffusion, the lipids will encounter restrictions in their translational motion. We choose to look at the system of eSM/CHOL that exhibits domain formation in the range of 6–22 mol% CHOL, and this is seen as a sudden break in the curve D_L versus CHOL (Figure 6.33). Due to fast exchange between the separated phases, l_0 and l_d , the observed diffusion coefficient, $D_L(2\phi)$, will be a weighted average of the diffusion coefficients in the separate phases:

$$D_L(2\phi) = p_o D_L(l_o) + (1 - p_o) D_L(l_d)$$
(3)

in which p_o is the relative amount of the l_o phase. As p_o increases the diffusion will decrease as $D_L(l_o)$ becomes dominant (note: lower diffusion because of closer packing in the l_o phase). A fast exchange is an indication that the domains are smaller than the mean diffusion length (ca. 1 μ m) calculated from $r^2 = 4D_L t$. In the one-phase regions on each side of the two-phase area, the diffusion coefficients are almost constant. In Figure 6.32, the two-phase area with the fluid phases, l_o and l_d , are indicated by 2ϕ .

For a ternary system with saturated and unsaturated lipid and CHOL, larger domains are observed and there is no exchange between the domains in the NMR experiment (cf. Figure 6.32). Another interesting finding was that the lateral diffusion is the same for all components, independent of the molecular structure (including cholesterol), if they reside in the same domain or phase in the membrane.

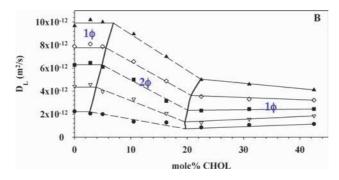


Fig. 6.33 Lateral eSM diffusion coefficients at different CHOL concentrations for the eSM/CHOL system with 35 wt% ²H₂O and at 313 K (circle), 318 K (triangle top down), 323 K (square), 328 K (diamond) and 333 K (triangle top up). The thick solid lines are estimations of the extension of the two-phase area (with l_0 and l_d phases). The solid and dotted lines are linear estimations of D_I in the one-phase, 1 ϕ and two-phase, 2 ϕ , areas, respectively. (Reprinted with permission from Filippov A, Orädd G, Lindblom G. (2003) The effect of cholesterol on the lateral diffusion of phospholipids in oriented bilayers. Biophys J 3079–3086. Copyright (2003) Biophysical Society.)

How can we get an understanding of the driving forces behind the domain formation in lipid bilayers? The lateral phase separation or domain formation in the ternary system with low- $T_{\rm m}$ and high- $T_{\rm m}$ lipids (for example, DOPC and eSM) can be rationalized in terms of lipid order and miscibility of unsaturated lipids in ordered phases. High- $T_{\rm m}$ lipids, such as DPPC and eSM, form more ordered phases than low- $T_{\rm m}$ analogs, such as DOPC and SDPC, and addition of CHOL greatly enhances the ordering, especially for the high- $T_{\rm m}$ lipids. Thus, it has been proposed that the lateral phase separation into l_d and l_o phases is entropy driven. This originates from the increasing difficulty to incorporate an unsaturated lipid with its bulky bending double bonds into a highly ordered phase, consisting of mainly saturated lipid and cholesterol. An unsaturated lipid prefers to be in an l_d phase, while a saturated lipid prefers to reside in the l_o phase (see Further Reading).

The plasma membrane caveolae (from the Latin for "little cavities) constitute a lipid raft subtype (Figure 6.34). They typically appear as microscopic, flask-shaped invaginations along the membrane surface of endothelial cells, adipocytes (fat cells) and smooth muscle cells. The principal protein component of caveolae is caveolin, a scaffolding protein that binds cholesterol efficiently and interacts with various signaling macromolecules, including G-proteins and calcium-regulating proteins. Caveolin may also regulate intracellular and surface cholesterol levels. Experiments with knockout mice lacking the caveolin-1 protein, and thus caveolae, demonstrated marked defects in arterial relaxation, myogenic tone, and exercise tolerance as a result of abnormalities in cell signaling and NO metabolism.

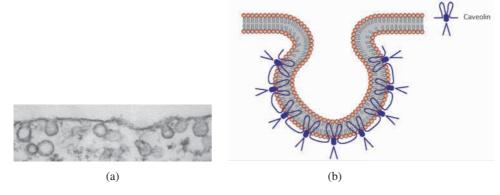


Fig. 6.34 ■ (a) Thin-section electron microscopy (EM) image and (b) cartoon of the interaction of the protein caveolin with the lipid bilayer.

Lipid rafts have also been suggested to play a role in diseases including atherosclerosis, hypertension, Alzheimer's disease, prion disease and viral infection.

6.4 Lipoproteins — "Good" and "Bad" Cholesterol

Lipoproteins are important protein-lipid assemblies that are responsible for the transport of fats to different parts of the body via the bloodstream. There are five major groups of lipoproteins, and all of them have an interior core that is composed of cholesteryl esters and triacylglycerols. These two types of molecules are hydrophobic and are therefore only slightly soluble in aqueous solutions such as the bloodstream. Lipoproteins solve this problem by coating the hydrophobic interior with an amphiphilic layer of phospholipids and unesterified cholesterol, i.e. the lipoprotein aggregate is held together by noncovalent forces (cf. micelles and lipid bilayers). The final components of lipoproteins are proteins called apolipoproteins. At least nine different apolipoproteins associate with human lipoproteins in substantial amounts, but the structures of most of them are similar. The structure consists of a high alpha helix content. One side of the helix tends to contain non-polar residues, while the other contains polar residues. The non-polar residues associate with the non-polar tails of the phospholipids in the lipoprotein and the polar residues associate with the polar head groups. Thus, the apolipoproteins surround the lipids to make up the lipoproteins (see Figure 6.35). LDL exhibits a thermal liquid crystalline-toisotropic transition of its cholesteryl esters between 25°C and 35°C. Quite recently, a threedimensional structure of LDL and LDLr complex was reported, where electron cryomicroscopy was used (Figure 6.35).

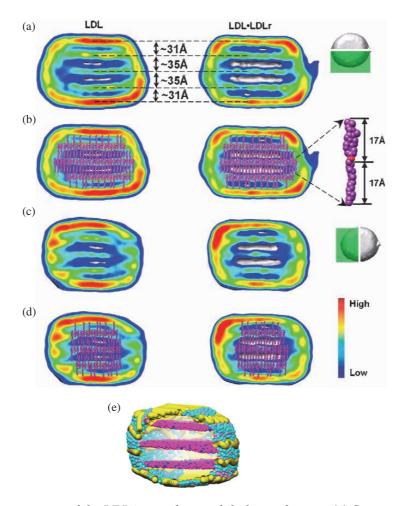


Fig. 6.35 ■ The structure of the LDL internal core of cholesteryl esters. (a) Cut-away surface views of the three-dimensional density maps of LDL (left) and the LDL. LDLr complex (right). Both cores contain striations separated by about 35Å at the center of the particle and about 31Å near the surface. (b) The cores largely comprise cholesteryl esters, which are modeled as juxtaposed stacks. In this model, the high-density sterol moieties of the cholesteryl ester molecules (magenta) are coplanar and their acyl chains extend outwards on either side into parallel planes that form the four lower density compartments. (c) and (d) show corresponding views perpendicular to (a) and (b), respectively. Similar features are seen including the low-density gaps demonstrating that the internal striations span the whole core, and accommodate cholesteryl ester in coplanar layers. (e) A cut-away view of the LDL model shows the surface structure of apo B-100 and the internal organization of the cholesteryl ester sterol moieties. The phospholipid head groups, cholesteryl esters and triacylglycerol are displayed as cyan, magenta and blue balls, respectively. According to the core dimensions observed and the partial specific volume of the cholesteryl ester, the striated core can accommodate about 1200 cholesteryl ester molecules. Calculations based on LDL composition give a similar value of about 1400. The additive length of the longest dimension of cholesterol (ca. 17Å) and the 18-carbon acyl chain (17Å) is ca. 34Å, which is consistent with the

Fig. 6.35 ■ (Continued) X-ray scattering measurements. The dimensions of cholesteryl ester are compatible with the juxtaposed stacking model of cholesteryl esters, with the sterol moieties spaced at 34 Å. The smaller spacing (31 Å) seen near the periphery is likely due to the protein elements of the outer shell, for example, the 8Å-thick betasheet-rich domain, being thinner than the 17 Å sterol moiety and because the smaller number of acyl chains in the outer stack of sterol moieties permits some chain tilting. (Reprinted with permission from Ren G, Rudenko G, Ludtke SJ, et al. (2010) Model of human low-density lipoprotein and bound receptor based on cryoEM. PNAS 107: 1059–1064. Copyright (2010) National Academy of Sciences USA.)

Lipoproteins are classified by their density. Since the protein components are denser than the lipids, lipoproteins with smaller percentage of protein are lower in density. The different groups of lipoproteins are very low-density lipoprotein (VLDL), low-density lipoprotein (LDL), intermediate-density lipoprotein (IDL), high-density lipoprotein (HDL) and chylomicrons. LDL and HDL are of particular interest due to their impact on human health - higher levels of LDL is believed to promote health problems and cardiovascular disease, while those with higher levels of HDL seems to correlate with a lower risk of cardiovascular disease. Therefore LDL is commonly referred to as "bad" cholesterol, while HDL is referred to as "good" or healthy cholesterol.

6.4.1 Apolipoproteins and Lipid Bilayers Can Form **Nanodiscs**

Structural studies of integral membrane proteins are hampered by the difficulties of finding appropriate membrane-mimicking media that maintain the protein structure and function. Typically detergent micelles are used to extract proteins out of the bilayer, and usually such micellar systems do not serve a natural habitat for a membrane protein to function properly. A relatively recent membrane system, called phospholipid nanodiscs, consisting of a lipid bilayer surrounded by two copies of lipid-binding protein, has become a popular system in the membrane protein field. In this system, membrane proteins can be investigated in a real detergent-free and nativelike lipid environment. Smaller nanodiscs tailored to accommodate membrane proteins of different sizes facilitate the structure determination of membrane proteins with NMR spectroscopy. Three-dimensional structural information about membrane proteins in an almost natural environment is obtained by NMR spectroscopy, and the use of nanodiscs also enables determination of protein dynamics in a lipid bilayer (see Figure 6.36). The most commonly used nanodisc has a diameter of 10 nm and a thickness of about 4 nm.

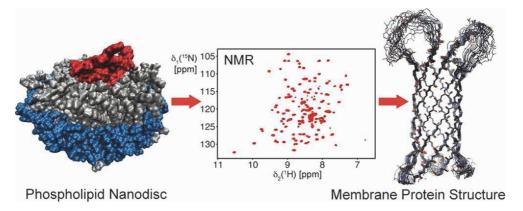


Fig. 6.36 ■ Size-optimized phospholipid nanodiscs for structural studies of membrane proteins with NMR spectroscopy. Nanodiscs consist of lipids encircled by two copies of apolipoprotein A-1 (ApoA-1). The length of the ApoA-1 defines the diameter of the nanodisc. Truncated ApoA-1 constructs lead to smaller nanodiscs, which are suitable for high-resolution structure determination by NMR spectroscopy. (Reprinted with permission from Hagn F, Etzkorn M, Raschle T, Wagner G. (2013) Optimized phospholipid bilayer nanodiscs facilitate high-resolution structure determination of membrane proteins. J Am Chem Soc 135: 1919–1925. Copyright (2013) American Chemical Society.)

For Further Reading

Nomenclature and Web-Sites

Fahy E et al. (2005)A comprehensive classification system for lipids. J Lipid Res 46: 839–861.

LIPID MAPS, http://www.lipidmaps.org; http://lipidlibrary.co.uk; http://lipidbank.jp; http:// www.lipidat.chemistry.ohio-state.edu and http://www.cyberlipid.org

International Union of Pure and Applied Chemists, and the International Union of Biochemistry and Molecular Biology (IUPAC-IUBMB) (see further reading for URL address http://www. chem.qmul.ac.uk/iupac/

Christie WW. What is a lipid? http://www.lipidlibrary.co.uk

Original Articles

Alley SH, Ces O, Templer RH, Barahona M. (2008) Biophysical regulation of lipid biosynthesis in the plasma membrane. *Biophys J* **94:** 2938–2954.

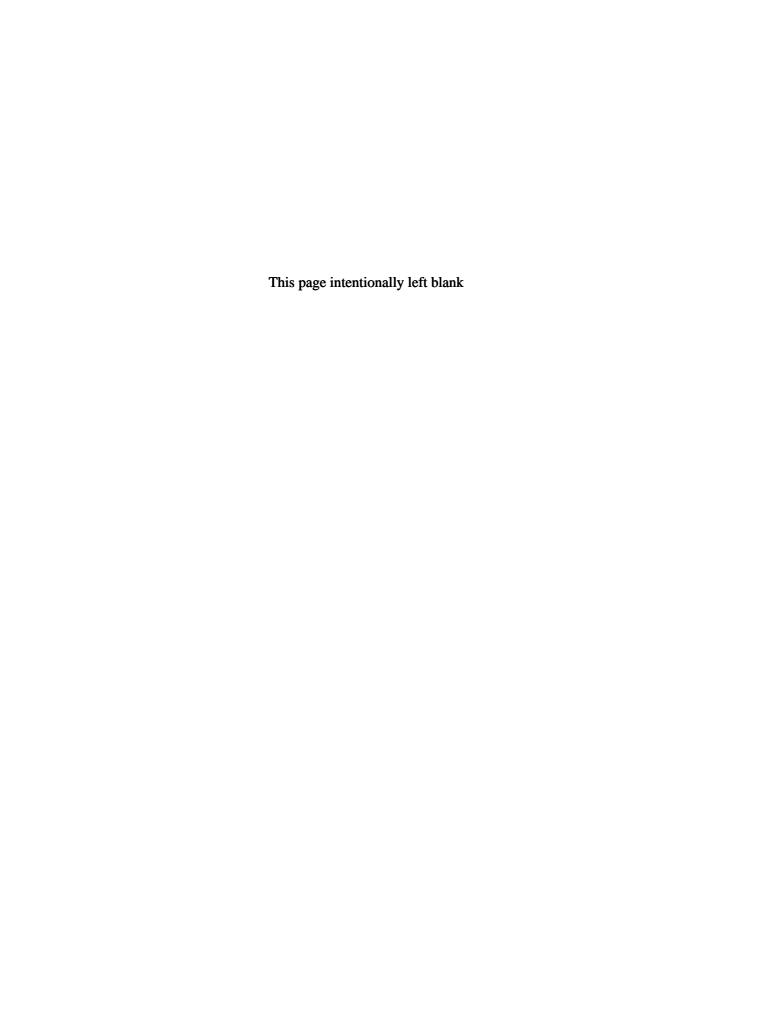
Gibson NJ, Brown MF. (1993) Lipid headgroup and acyl chain composition modulate the MI-MII equilibrium of rhodopsin in recombinant membranes. *Biochemistry* **32**: 2438–2454.

- Ipsen JH, Karlström G, Mouritsen OG, et al. (1987) Phase equilibria in the phosphatidylcholine-cholesterol system. *Biochim Biophys Acta* **905**:162–172.
- Killian JA, Salemink I, de Planque MRR, *et al.* (1996) Induction of nonbilayer structures in diacylphosphatidylcholine model membranes by transmembrane α-helical peptides: Importance of hydrophobic mismatch and proposed role of tryptophans. *Biochemistry* **35**: 1037–1045.
- Lindblom G, Brentel I, Sjölund M, *et al.* (1986) Phase equilibria of membrane lipids from *Acholeplasma laidlawii*. The importance of a single lipid forming nonlamellar phases. *Biochemistry* **25**: 7502–7510.
- Morein S, Andersson A-S, Rilfors L, Lindblom G. (1996) Wild-type *Escherichia coli* cells regulate the membrane lipid composition in a "window" between gel and non-lamellar structures. *J Biol Chem* **271**: 6801–6809.
- Rilfors L, Lindblom G. (2002) Regulation of lipid composition in biological membranes Biophysical studies of lipids and lipid synthesizing enzymes. *Colloids Surf B Biointerfaces* **26**: 112–124.
- Sparr E, Åberg C, Nilsson P, Wennerström H. (2009) Diffusional transport in responding lipid membranes. *Soft Matter* **25**: 3225–3233.
- Veatch SL, Keller SL. (2005) Seeing spots: Complex phase behavior in simple membranes, *Biochim Biophys Acta* **1746**: 172–185.
- Vist MR, Davis JH. (1990) Phase equilibria of cholesterol/dipalmitoylphosphatidylcholinemixtures: ²H nuclear magnetic resonance and differential scanning calorimetry. *Biochemistry* **29:** 451–464.

Reviews and Books

- Brown MF. (2012). Curvature forces in membrane-protein interactions. *Biochemistry* **51:** 9782–9795. Chernomordik LV, Zimmerbergh J, Kozlov MM. (2006) Membranes of the world unite! *J Cell Biol* **175:** 201–207.
- Nicolson GL. (2014) The fluid-mosaic model of membrane structure: Still relevant to understanding the structure, function and dynamics of biological membranes after more than 40 years. *Biochim Biophys Acta* **1838**: 1451–1466.
- Killian JA. (2003) Synthetic peptides as models for intrinsic membrane proteins. *FEBS Lett* **555**: 134–138.
- Lindblom G. (1996) NMR spectroscopy on lipid phase behaviour and lipid diffusion. In *Advances in Lipid Methodology*, WW Christe (ed.), Oily Press, Ltd., Dundee, Scotland, pp. 133–209.
- Lindblom G, Orädd G. (2009) Lipid lateral diffusion and membrane heterogeneity. *Biochim Biophys Acta* **1788**: 234–244.
- Luckey M. (2011) *Membrane Structural Biology*. Cambridge University Press, New York, NY, USA, ISBN 978-0-521-85655-3.
- Marsh D. (2009) Cholesterol-induced fluid membrane domains: A compendium of lipid-raft ternary phase diagrams. *Biochim Biophys Acta* **1788**: 2114–2123.

- McIntosh TJ (ed.) (2007) Lipid rafts. Meth Mol Biol, Vol. 398, Springer Verlag, Berlin.
- Mouritsen O. (2005) Life As a matter of fat. The emerging science of lipidomics. Springer Verlag, Berlin, Heidelberg GmbH & Co. K., ISBN 3-540-5 23248-6.
- Owen DM, Gaus K. (2013) Imaging lipid domains in cell membranes: The advent of superresolution fluorescence microscopy. Front Plant Sci 4: 1–9.
- Owen DM, Magenau A, Williamson D, Gaus K. (2012) The lipid raft hypothesis revisited New insights on raft composition and function from super-resolution fluorescence microscopy. *Bioessays* **34**: 739–747.
- Sherman GC, Tyler A II, Brooks NJ, et al. (2010) Ordered micellar and inverse micellar lyotropic phases. Liq Cryst 37: 679-694.
- Tanford C. (1980) The hydrophobic effect. Wiley, New York, USA.
- Yeagle PL. (2005) The structure of biological membranes. CRC Press, Boca Raton, Florida, USA, ISBN 0-8493-1403-8.



Basics of Carbohydrates

Carbohydrates or sugars, are very abundant in nature. Most have the chemical formula $C_m(H_2O)_n$. Thus they could be said to be hydrates of carbon. Carbohydrates occur as single units (monosaccharides), short chains of linked monosaccharide units (oligosaccharides), or extended chains (polysaccharides). Carbohydrates are frequently used for energy storage in the form of starch in plants or glycogen in animals. However, carbohydrates also have important structural and functional roles (Figure 7.1).

Cellulose is the most abundant organic polymer on earth (Table 7.1). Green plants produce cellulose to form their cell wall, but many other species also form cellulose. Cotton fiber is extremely high in cellulose and so is wood. A relative of cellulose is chitin, which builds up the cell walls of fungi and the exoskeletons of arthropods (crabs and shrimps) and the wings of insects. In addition, there are extracellular carbohydrates, like hyaluronan, that have important functions as a transport barrier in the extracellular matrix, but also as a signal molecule in developmental processes and in defense mechanisms.

Bacteria, particularly gram-positive bacteria, have an extensive peptidoglycan structure on the outside of the plasma membrane to form the structural cell wall. The peptidoglycan is composed of long chains of two alternating kinds of monosaccharide units crosslinked by short peptides.

Yet another large group of carbohydrates in nature are the glycoproteins. More than 50% of all human proteins are glycosylated. One example of glycosylated proteins are the antibodies. A forest of carbohydrates bound to proteins or lipids cover many eukaryotic cells and viruses, like HIV or influenza virus that have several heavily glycosylated proteins on their surface. A class of proteins, called lectins, binds specifically to different carbohydrates (Figure 7.1).

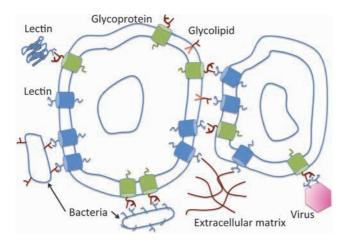


Fig. 7.1 ■ A schematic illustration of some of the interplay with cells where carbohydrates are involved. Carbohydrates are illustrated in brown, glycoproteins in green and lectin proteins, either membrane bound or free in solution, in blue.

| Name | Species | Location | Role |
|---------------|------------|----------------------------|--|
| Cellulose | Plant | Cell wall | Protective |
| Hemicellulose | Plant | Cell wall | Protective |
| Chitin | Fungi, | Cell wall | Protective |
| | Arthropods | Exoskeleton | Protective |
| Peptidoglycan | Bacteria | Cell wall | Protective |
| Hyaluronan | Eukaryotes | Extracellular matrix | Lubricant, cell adhesion, cell signaling |
| Proteoglycans | Eukaryotes | Membrane bound or secreted | Scaffold functions, cell signaling |
| Glycolipids | All | Membrane bound | Cellular recognition? |
| Glycoproteins | All | Membrane bound or secreted | |

7.1 The Common Monosaccharide Units

A very large number of monosaccharides types occur in nature, at least 12 in eukaryotes. Bacteria have a much wider range of monosaccharides. We will primarily limit our presentation to the eukaryotic ones (Table 7.2).

Character Name # Carbons Examples Pentose 5 Neutral D-xylose [Xyl], ribose, deoxyribose Hexose 6 Neutral D-glucose [Glc], D-galactose [Gal] D-mannose [Man] Hexosamine 6 Amino group at 2-position, N-acetyl-D-glucosamine free, acetylated or [GlcNAc], sulfated N-acetyl-D-galactosamine [GalNAc] Deoxyhexose Lacking hydroxyl group at L-fucose [Fuc] 6 6-position Uronic acid 6 Carboxylate at 6-position D-glucuronic acid [GlcA] L-iduronic acid [IdoA] Sialic acid 9 Acidic N-acetylneuramic acid [Neu5Ac]

TABLE 7.2 Common Types of Monosaccharides in Higher Animals

Glucose is the central monosaccharide, which can be converted into other sugars. Figure 7.2 shows glucose in both a linear and cyclic form with 6 (pyranose) or 5 atoms (furanose) in the ring. The cyclic forms have two major forms (α and β), depending on the orientation of the OH group next to the oxygen in the ring. Figure 7.3 shows a number of common cyclic monosaccharides. The monosaccharides primarily occur in their cyclic form. Linear monosaccharides contain at least one asymmetric carbon atom. The number of asymmetric carbon atoms equals the number of internal CHOH groups. Aldohexose (general formula $C_6H_{12}O_6$) has four asymmetric carbon atoms and 16 different isomeric forms. The L- and D-versions of glucose are two of the 16 forms (Figure 7.3, bottom). The conformation of the asymmetric carbon furthest from the aldehyde carbon defines the D or L prefix. In hexoses, this is C-5 and in pentoses it is C-4. When linear monosaccharides cyclize yet another asymmetric carbon is introduced, which is identified by α or β . In addition, the hydroxyl groups can be modified by deoxygenation (changing a hydroxyl to a hydrogen), phosphorylation, sulfation, methylation, O-acetylation or addition of N-acetamido groups or fatty acids.

The planar conformations are simplified representations (Figures 7.2 and 7.3, top). Monosaccharides normally adopt a chair conformation (Figure 7.3, bottom). Several less stable variants occur, like planar, half-chair, boat and twist-boat.

Sugar moieties have a displacement of its electrons in such a way that the hydroxyl groups on one side of the sugar ring draw electrons and generate a partial positive charge on the hydroxyl free surface of the sugar moiety. This partial positive surface preferably interacts with aromatic groups in proteins (see Section 2.2.3) where a charge-charge interaction can be established (Figure 7.4).

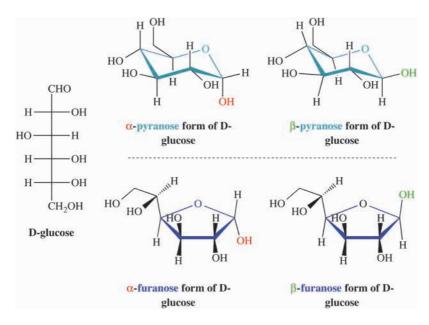


Fig. 7.2 \blacksquare A linear representation of the common sugar D-glucose and cyclic α and β forms of D-glucose, pyranose and furanose. L-glucose is the mirror image of D-glucose and is much less abundant. (Illustration prepared by Lars Erik Andreas Ehnbom.)

7.2 Glycosidic Bond Formation

Two monosaccharides can be joined by a glycosidic bond between the anomeric carbon atoms (the carbon atom of the aldehyde group in the linear form) in one monosaccharide with a hydroxyl group in another. A simple alcohol or a hydroxylated amino acid can also provide the hydroxyl group to form a glycosidic bond. Most glycosidic bonds can be hydrolyzed in dilute acids. However, the half-life for spontaneous hydrolysis at neutral pH and room temperature of starch and cellulose is in the order of millions of years.

An oligosaccharide has one non-reducing end and one reducing end. The latter is where the anomeric carbon is free for further reactions. The order of the units in an oligosaccharide is described from the non-reducing end.

7.2.1 The Glycosyl Transferases and Glycosyl Hydrolases

The enzymes involved in synthesizing carbohydrates or adding carbohydrates to proteins are called glycosyl transferases (GT). Three different protein folds have been identified, GT-A, GT-B and GT-C where the last one is a membrane bound enzyme engaged in

Fig. 7.3 ■ Some types of sugar residues shown in both Fisher projections and in chair conformations. (Illustration prepared by Lars Erik Andreas Ehnbom.)

cell wall synthesis. All three types are metal enzymes. There are also glycoside hydrolases removing part or the whole of the added carbohydrate. The glycosyl transferases account for 1–3% of all proteins encoded in the known genomes. While humans have about 230 glycosyl transferase genes, the genus *Populus* (poplars) has more than 800. Numerous structures of carbohydrate-handling enzymes are known. At the moment more than 200 families of carbohydrate-handling enzymes are classified (http://afmb.cnrs-mrs.fr/CAZY/). The enzymes have been described as belonging to four classes, glycoside

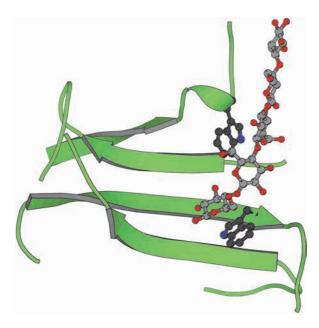


Fig. 7.4 ■ Interaction between aromatic groups and carbohydrates. Two tryptophan residues in the cellobiohydrolase CBH1 are part of the surface of a tunnel to which two D-glucose units of cellopentaose (grey carbon atoms) can bind (PDB: 6CEL).

hydrolases (135 families), glycosyl transferases (97 families), polysaccharide lyases (23 families) and carbohydrate esterases (16 families).

7.3 Glycoproteins and Glycolipids

Glycosylation is a covalent modification of a protein or a fatty acid by the addition of one or several carbohydrate moieties. Glycolipids are briefly discussed in Chapter 6. For glycosylated proteins the terms glycoprotein and proteoglycan are used.

Glycosylation can be enzymatic or non-enzymatic (glycation). Glycation may impair the function of the protein. Enzymatic glycosylation is performed at specific sites in the protein. As many as half of all proteins may be glycosylated, which makes it the most common post-translational modification. Glycosylation occurs in eukaryotes, archaea and also bacteria. Almost all secreted or membrane-bound eukaryotic proteins are glycosylated, whereas most cytosolic or nuclear proteins are not. The subject of glycosylation is very extensive, but probably still in its infancy and only a brief summary of the basics can be provided here.

The addition of sugar moieties to proteins is a complex process. Unlike protein synthesis, the addition of sugar moieties is not directly encoded in the genome. First, a large number of different kinds of sugar moieties can be added. Second, the glycosylation can be in the form of a single unit or several linked units with or without branches. Third, all hydroxyls of the sugar can be used for linkage. In estimating the way three specific monosaccharides could be combined, the result is several thousand different possible trisaccharides. The complexity increases dramatically if additional sugar molecules are part of the trisaccharide or if more than three sugar molecules form the glycosylation. Furthermore, a protein may have one or several amino acid residues that are modified by carbohydrates. The number of possible combinations becomes astronomical. Fortunately, the combinations that occur in nature are limited to a number of main variants.

The roles of the added oligosaccharides to proteins are manifold. In some cases oligosaccharides assist the proper folding of proteins. In other cases they can stabilize a protein and protect it from proteolysis. In yet other cases they participate in cell-cell contacts. More specific functions have been identified in recent years. Thus, a large number of glycosyl transferases modify epidermal growth factors. The O-GlcNAcylation of thousands of protein substrates by a transferase (OGT) links nutrient metabolism to gene expression. Specific glucosaminidases (OGA) can remove the monosaccharide. The level of this modification controls the numerous substrates, among them the transcription machinery including rRNA polymerase II and many transcription factors (see Chapter 10).

The enzymatic transfer of a monosaccharide to acceptors (e.g. monosaccharides, oligosaccharides, lipids, nucleic acids, antibiotics or proteins) uses an activated substrate as donor where the sugar moiety is linked to a phosphate of UDP, GDP, CMP or lipid phosphates. UDP is the dominating donor. The transfer normally occurs by one monosaccharide at a time. In the case of synthesizing an oligosaccharide the product of one enzyme becomes the substrate for the next. These enzymes are generally integral membrane proteins.

Proteins can be O- or N- glycosylated. O-glycosylation primarily occurs on the side chain oxygens of serine or threonine residues. Tyrosine side chains are less common to be O-glycosylated. This glycosylation probably occurs in the Golgi apparatus in eukaryotes. No special consensus sequence motif (or sequon) has yet been identified for O-glycosylations. O-linked oligosaccharides are generally short (1-4 sugar residues) and linked to the acceptor hydroxyl group through N-acetylgalactosamine. The donor is normally UDP. The different glycosyl transferases involved in these additions have well identified, but varied specificity for the specific amino acid sequences of the substrates. These enzymes are specific both for the sugar nucleotide molecule and the acceptor, being either a protein or a previously added sugar. Furthermore, the enzymes are specific for which hydroxyl group the new sugar will be added to.

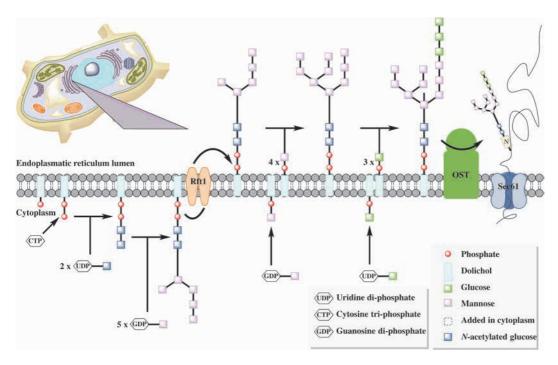


Fig. 7.5 ■ An example of how a precursor glycan is synthesized on the cytoplasmic side of the endoplasmic reticulum (ER) and flipped into the lumen of ER where it is completed. The oligosaccharide is subsequently transferred by the oligosaccharide transferase (OST) to the Asn residue of a growing polypeptide. (Illustration prepared by Lars Erik Andreas Ehnbom.)

N-glycosylation is the most common type of glycosylation and often occurs in the endoplasmic reticulum (ER) while the protein is being translated. ER is the site of synthesis of membrane proteins as well as those that are secreted. N-glycosylation takes place on the luminal surface of ER. In eukaryotes mono- or en bloc oligosaccharides are transferred to the side chain nitrogen of asparagines that are part of the consensus sequence or sequon, N-X-S/T where X cannot be proline (Figure 7.5). However, not all sequences with this motif are modified. In bacteria, the sequon is extended to D/E-X-N-X-S/T. The short eukaryotic recognition sequence makes the specificity broad. The modification involves a large branched oligosaccharide, which is prefabricated and attached to dolichol, a long lipid that is firmly bound in the membrane. Frequently, the oligosaccharide has the composition (glucose)₃(mannose)₉(N-acetylglucose)₂. After the transfer the oligosaccharide is trimmed by removing glucoses. The process is part of the checking mechanism for proper protein folding. Properly folded proteins are transported to the Golgi for further processing.

The glycosyl modifications are highly hydrophilic and therefore exposed on the surface of the proteins. Furthermore, the carbohydrate moieties tend to be highly flexible. For crystallographic work on proteins it is often favorable to remove the oligosaccharides

from the surface of the protein. Little or no conformational changes have been observed due to this removal.

7.4 Important Polysaccharides

A number of physiologically important polysaccharides are known. To these belong among others the homopolysaccharides cellulose and chitin, which both are composed of a single type of monosaccharide unit. The glycosaminoglycans, hyaluronan, chondroitin sulfate, dermatan sulfate, heparin and heparan sulfate, are all heteropolysaccharides composed of two alternating types of monosaccharide units (hexuronic acid and hexosamine). The sulfated glycosaminoglycans occur covalently bound to proteins in proteoglycan structures. Information is gradually accumulating concerning the structural biology of polysaccharide synthesis and degradation. Particularly in the case of cellulose, some knowledge has been gained on both sides.

7.4.1 Cellulose

7.4.1.1 Synthesis

The most common organic polymer on earth is cellulose. It is a major structural component of the primary cell wall of green plants. The amount on earth has been estimated to 7×10^{11} tons. Cellulose is a prime component of biomass, which can be used to generate bioenergy through a range of processes. A prime way is enzymatic hydrolysis to fermentable sugar molecules. Cellulose is a long linear chain of linked D-glucose units synthesized by cellulose synthase (Figure 7.6). The linkage is β -1,4 with 180° flips between neighboring glucose units.

In plants, cellulose polymers have about 36 parallel chains forming microfibrils with lengths of about 500 to 15 000 residues (Figure 7.7). These microfibrils are arranged into

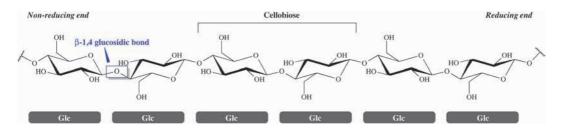


Fig. 7.6 ■ The structure of cellulose built of β -D-glucose units. (Illustration prepared by Lars Erik Andreas Ehnbom.)

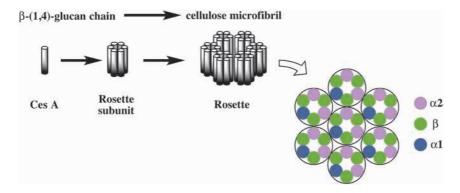


Fig. 7.7 ■ The arrangement of three types of subunits in the rosette of cellulose synthase that produces cellulose microfibrils (Doblin *et al.*, 2002). (Illustration prepared by Lars Erik Andreas Ehnbom.)

macrofibrils. Extensive hydrogen bonding is formed between the polymers, resulting in fibers of significant strength. Cellulose is synthesized from uridine diphosphate-glucose (UDP-glucose, the activated, donor form of glucose) by a multi-subunit enzyme complex in the plasma membrane composed of about 36 units providing the components needed to produce a microfibril. Cellulose comprises 30–35% of wood, while other polymers like hemicellulose and lignin each comprise 20–30% of its dry weight.

Bacteria also synthesize cellulose, partly for the formation of biofilms.

Starch is another homopolysaccharide composed of glucose units. These are joined by α -1,4 linkages, instead of the ß linkages found in cellulose. Because of this difference starch can be metabolized and used for energy generation by animals.

7.4.2 Chitin

Chitin is a long-chain polysaccharide composed of N-acetyl-D-glucosamine units that builds up the cell walls of fungi and the exoskeletons of arthropods (including crabs, shrimps and insects). It is the second most abundant polysaccharide in nature. It is closely related to cellulose, but one of the hydroxyl groups of each glucose residue of cellulose is exchanged for an acetylamine group.

The structural insight into the synthesis of chitin is very limited, whereas chitinases are more extensively studied. The breakdown products have a range of industrial and medical uses. Chitinases are found in viruses, bacteria, higher plants and animals. At least two different families are known with very different structures and catalytic mechanisms. One family, with the TIM-barrel fold, has a mechanism with retention of the anomeric conformation. Others have structures like the hen egg-white lysozyme with a high α -helical content and an inverting mechanism.

7.4.3 Hyaluronan

The view of polysaccharides has grown by the identification of a number of important physiological roles for these molecules. Hyaluronan (HA, Figure 7.8) is a large linear glycosaminoglycan. It has molecular weights up to 10^7 and is highly negatively charged and also heavily hydrated. It is found in the extracellular matrix, in the synovial fluid, umbilical cord and in the vitreous body of the eye. It functions as a lubricant in joints and in cartilage it interacts with chondroitin sulfate proteoglycan to form macromolecular complexes that provide resistance to compression. It has roles in cell adhesion, apoptosis, migration and proliferation. The synthesis of HA is tightly controlled. In mammals, it is synthesized by three isoenzymes in the plasma membrane. A range of different classes of proteins can bind to HA.

Various hyaluronidases cleave HA and the products range in size from disaccharides to fragments of 50–60 disaccharide units (molecular weights around 20 000). These oligosaccharides can participate in cell-cell signaling.

7.4.4 Heparin/Heparan Sulfate

Heparin and heparan sulfate are related and important carbohydrates involved in a wide range of biological processes. A simple disaccharide repeat of uronic acid and glucosamine is extensively modified with O- and N-sulfo and N-acetyl groups. The overall molecular weight ranges from 5 to 40 kDa. Heparin, which is confined to mast cells, is the most negatively charged biological molecule. Mast-cell heparin occurs in proteoglycan form, tightly bound to proteases, but can be fragmented by an endoglucuronidase and released into the extracellular matrix. Whereas heparin can interact with a range of different proteins, its overall physiological role remains unclear. Heparan sulfate is generally less sulfated, in a cell-specific and highly variable manner. It forms part of the extracellular matrix and is also abundant at the cell surface as part of cell surface receptors.

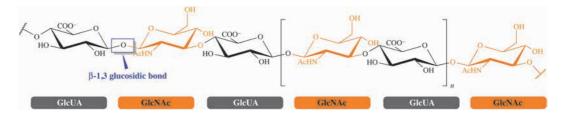


Fig. 7.8 ■ Hyaluronan (HA) has a disaccharide repeat (glucuronic acid (GlcUA) and N-acetylglucosamine GlcNAc)) of on average 10 000 units. The connection is by a β -1,3 glycosidic bond. (Illustration prepared by Lars Erik Andreas Ehnhom.)

7.4.4.1 Heparin in blood coagulation

Heparin activates the serine protease inhibitor antithrombin (serpins, see Section 12.3), which is a central inhibitor of several of the proteases involved in blood coagulation. Thrombin is the final protease of the blood coagulation cascade responsible for the cleavage of fibrin to form the fibrin clot. Antithrombin is part of the control of the activity of thrombin. Antithrombin is activated by heparin, but potentially also by heparan sulfate in the vascular system, due to interaction with a specific pentasaccharide sequence. While the functional role of (the extravascular) anticoagulant mast-cell heparin remains unclear, heparin is nevertheless one of the most frequently used natural drugs, clinically applied to prevent thrombosis.

Heparin can also bind to fibroblast growth factors with high affinity and assists in their interaction with the fibroblast growth factor receptor. This receptor is built on several Ig domains (see Section 14.1).

7.5 Cell Walls and Extracellular Matrix

In many instances, cells have a complex outer layer, a cell wall, where carbohydrates have a significant role. In bacteria, the cell wall is called peptidoglycan or murein and is a network outside the cytoplasmic membrane and composed of alternating N-acetylglucosamine and N-acetylmuramic acid residues crosslinked by short peptides. In plants, cell walls are formed primarily of cellulose, hemicellulose and pectin. These cell walls can be several μm thick in order to withstand a strong internal osmotic pressure. Animal cells do not have a cell wall.

Multicellular structures normally have an extracellular matrix (ECM) with common functions that concern cell adhesion, cell-to-cell communication and differentiation (see Chapter 16). The ECM is composed of an interlocking mesh of fibrous proteins and glycosaminoglycans (GAGs). Many of the GAG polymers (heparan sulfate, chondroitin/dermatan sulfate and keratan sulfate) are attached to core proteins of proteoglycans, but hyaluronan is an exception. The main protein in ECM is collagen, but elastin is also an important component making the tissues elastic. Fibronectin is a glycoprotein that connects cells with the collagen fibers and assists in cellular movements and reorganization. Integrins are protein receptors connecting the inside of cells with other cells or molecules in the ECM.

Web Sites

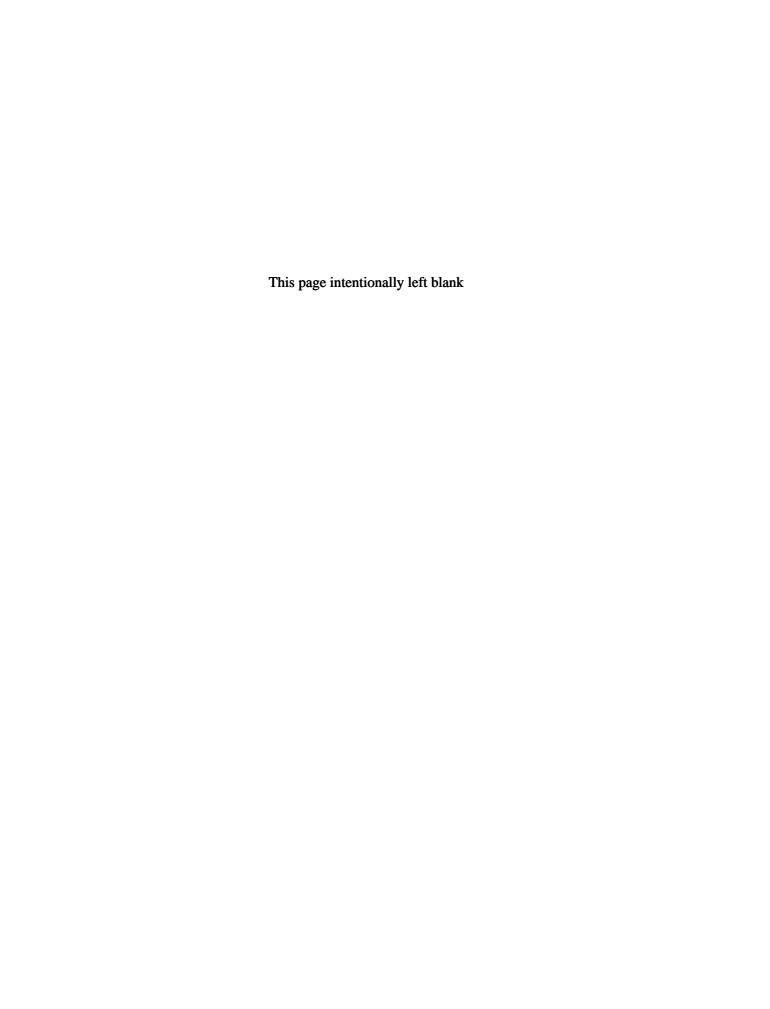
Further Reading

Original Articles

- Li W, Johnson DJKD, Esmon CT, Huntington JA. (2004) Structure of the antithrombin-thrombin-heparin ternary complex reveals the antithrombotic mechanism of heparin. *Nat Struct Mol Biol* 11: 857–862.
- Lizak C, Gerber S, Numao S, et al. (2011) X-ray structure of a bacterial oligosaccharyltransferase. *Nature* **474**: 350–355.
- Morgan JLW, Strumillo J, Zimmer J. (2013) Crystallographic snapshot of cellulose synthase and membrane translocation. *Nature* **493**: 181–187.
- Shaya D, Tocilj A, Li Y, et al. (2006) Crystal structure of heparinase II from *Pedobacter heparinus* and its complex with a disaccharide product. *J Biol Chem* **281**: 15525–15535.

Books and Reviews

- Arki *et al.* (2009) *Essentials of Glycobiology*. 2nd ed. Cold Spring Harbor Press. Cold Spring Harbor. New York.
- Boraston AB, Bolam DM, Gilbert HJ, Davies GD. (2004) Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. *Biochem J* **382**, 769–781.
- Breton C, Fournel-Gigleux S, Palcic MM. (2012) Recent structures, evolution and mechanisms of glycosyltransferases. *Curr Opin Struct Biol* **22**: 540–549.
- Hurtado-Guerrero R, Davies GJ. (2012) Recent structural and mechanistic insights into post-translational enzymatic glycosylation. *Curr Opin Chem Biol* **16**: 479–487.
- Lairson LL, Henrissat B, Davies GJ, Withers SG. (2008) Glycosyltransferases: Structures, functions and mechanisms. *Ann Rev Biochem* 77: 521–555.
- Malik V, Black GW. (2012) Structural, functional, and mutagenesis studies of UDP-glycosyl transferases. *Adv Prot Chem Struct Biol* **87**: 87–115.
- Sandgren M, Ståhlberg J, Mitchinson C. (2005) Structural and biochemical studies of GH family 12 cellulases: Improved thermal stability and ligand complexes. *Prog Biophys Mol Biol* **89**: 246–291.
- Wang M, Liu K, Dai L, *et al.* (2013) The structural and biochemical basis for cellulose biodegradation. *J Chem Technol Biotechnol* **88**: 491–500.



Enzymes

Enzymes are proteins that catalyze biochemical reactions without being consumed, and are able to perform the same reaction over and over again. They have a wide range of catalytic properties and are classified on that basis (hydrolases, ligases, reductases, oxidases and so on).

Enzymes are usually large molecules, but only a small fraction of the amino acid residues participate in the catalysis. The area of an enzyme where the binding of the substrate(s) and the catalysis occurs is called the active site. The active sites are frequently located in some sort of depression or cavity in the structure of the enzyme. Sometimes cofactors (like metal ions) or coenzymes (like NADH) are bound in the active site and participate in the reaction.

Many enzymes are highly specific for their substrates. This is generated by complementarities in shape of the substrate and the active site. The complementarity may also include the charge, polarity and hydrophobicity relationship between substrate and active site. Due to this complementarity enzymes are often highly stereospecific, substrates with the wrong hand may not be able to bind. Different models have been used to describe the interaction between enzyme and substrate. An early description is the "lock and key model" which illustrates the complementarity but not how the enzyme may function.

Enzymes are flexible molecules like all proteins. The dynamics involve atomic oscillations, side chain reorientation and movements of main chain or of whole domains. This dynamic character is essential for enzyme activity. During binding and catalysis residues of the active site or large parts of the enzymes can undergo conformational changes just like the substrate going through chemical changes. One model that emphasizes the conformational changes of enzyme as well as substrate is called "induced fit" (Figure 8.1). With a conformational change, the groups participating in catalysis get close to each other.

A general role of an enzyme is to lower the activation energy of the chemical reaction (Figure 8.2). This can simply be done if the enzyme binds the substrates with optimal

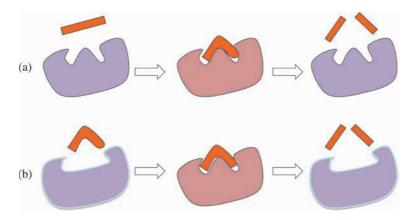


Fig. 8.1 ■ (a) Simplified enzyme mechanisms. (b) The substrate is forced to bind to the enzyme in a strained conformation, which leads to the degradation of the substrate. The enzyme undergoes a conformational change to bind the substrate and catalyze the reaction. In (a) and (b), the substrate or the enzyme, respectively undergoes an induced fit. The enzyme in the red state stabilizes the transition state.

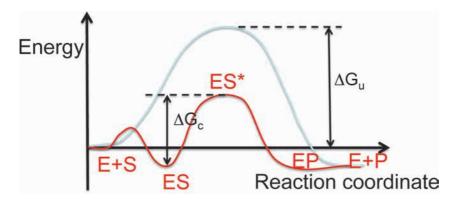


Fig. 8.2 ■ Chemical reactions have an activation energy. The uncatalyzed reaction is shown with the blue curve. Enzymes catalyze the reaction (red curve) by lowering the activation energy. E means enzyme, S — substrate and P — product. ES* indicates the activated enzyme substrate complex at the transition state.

proximity and orientation for the reaction to occur. This is analogous to an increase in concentration of the substrates or product.

The affinity of an enzyme for the transition state may be greater than for the substrate or the product ("transition state stabilization"). This causes a strain in the substrate. In enzymes that catalyze reactions where covalent bonds are formed or broken, the enzyme may make the substrates bind closer than van der Waals' distances to assist in the formation

of the bond that should be formed. In the opposite direction, the enzyme could strain or pull a covalent bond that should be broken. The uncatalyzed reaction obviously is devoid of this possibility.

Many enzymes catalyze reactions where protons are removed or added (acid/base catalysis). Residues of the enzymes, often histidines with a pKa close to neutral, participate in these processes. Frequently, pKa values of groups in active sites can be significantly shifted due to the electrostatic interactions.

Electrostatic effects are often of significant importance in catalysis to activate nucleophiles or electrophiles or to stabilize leaving groups. Here charged amino acids or metal cofactors have the most pronounced roles.

Some enzymes (e.g. serine proteases like trypsin) form an intermediate covalent complex with the substrate. This is obviously a reaction path that only can take place in the presence of the enzyme. Likewise aminoacyl-tRNA synthetases initially activate the amino acid with an ATP molecule before it is transferred to the tRNA substrate.

Some enzymes are allosteric and are regulated by the binding of small molecules. This binding can activate or inhibit the enzyme by inducing conformational changes that affects the active site. For many enzymes, it would be impossible to analyze the catalytic mechanisms without using inhibitors that can stop the enzyme in any of the conformational states it is going through.

In this chapter, we present a few enzymes with very different modes of action and important biological roles: an extremely rapid enzyme; a highly regulated enzyme; a family of enzymes that are molecular motors; a molecular switch and a multifunctional enzyme. Additional enzymes will be described in other chapters of this book.

8.1 Carbonic Anhydrase — An Extremely Rapid Enzyme

Carbonic anhydrase (CA) catalyzes a simple reaction, it hydrates carbon dioxide or dehydrates bicarbonate. The enzyme is found in all cells and exists in five major forms, α , β , γ , δ and ζ . For several of these forms there are many different isoenzymes with different physiological tasks. This is in particular true for the α -form. At least three of the forms of the enzyme have evolved independently from different origins, with different tertiary and quaternary structures. Part of the ζ -form must have a common origin with the β -form. The different forms catalyze the reaction in very similar ways (Figure 8.3). The interest in this enzyme is its catalytic rate and its multifaceted physiological roles. The maximal turnover rate in carbon dioxide hydration is $10^6 \, \text{s}^{-1}$ for an enzyme of the α -form, which is close to the diffusion limit. How is the enzyme designed, and what is

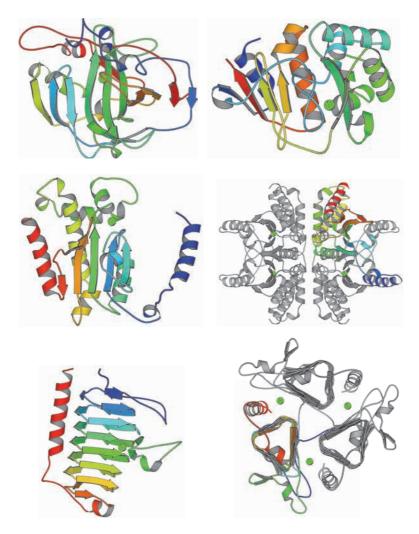


Fig. 8.3 \blacksquare Four different versions of carbonic anhydrases, α (top left, PDB: 2CBB), ζ (top right, PDB: 3BOB), β (*middle*, PDB: 1I6P) and γ (*bottom*, PDB: 1QRE). The folds are totally different. In addition, the α -form is monomeric while the β -form is tetrameric, hexameric or octameric, formed by two, three or four dimers. The γ -form of the enzyme is a trimer of β -helices. One subunit in each case is shown in color. The active site is indicated by the zinc ions (green). In the γ -form of the enzyme, the active site is formed at the subunit interface.

the mechanism that allows it to function at such a high rate? The steps of the reaction are the following:

$$Zn-H_2O = Zn-OH^- + H^+$$

 $H^+ + His = His - H^+$
 $His-H^+ + Buffer = His + Buffer - H^+$
 $Zn-OH^- + CO_2 + H_2O = HCO_3^- + Zn - H_2O$

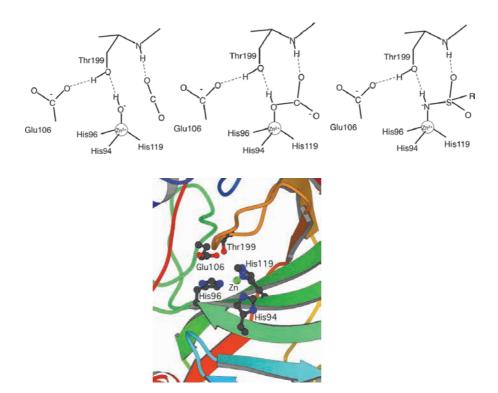


Fig. 8.4 ■ The active site of α -carbonic anhydrase. *Top left*: The hydroxyl is ready for nucleophilic attack on the carbon dioxide. *Top middle*: The substrate HCO_3^- is bound with the protonated oxygen but not the negatively charged oxygen to the zinc ion. This is due to the fact that the OH group of Thr199 is hydrogen bonded to Glu106. The position and orientation of the Thr199 OH is therefore such that it prevents non-protonated groups from binding at the zinc ion. The rapid catalysis is due only to the breakage or formation of the bond between the OH and carbon of the bicarbonate. *Top right*: The mode of binding of the strong sulfonamide inhibitors. The negatively charged NH-group is optimal as a fourth ligand position at the zinc ion with its proton is hydrogen bonded to Thr199. This type of inhibitor is a transition state analog. *Below*: A view of the active site of the enzyme (PDB: 2CBB).

A metal ion is the common component in the active sites of these enzymes, normally zinc. However, the γ -form probably has an iron ion *in vivo* and the ζ -form uses cadmium. Both these forms also function with zinc. In the α and γ enzymes, the metal ion is bound to three histidyl residues, but to a histidyl and two cysteinyl residues in the β and ζ classes. The charge of the metal ion environment is 2+ in the α - β -and δ -forms (Figure 8.4).

The metal ion is essential for activating the water molecule by lowering its pKa value to around 7. The active form of the enzyme thus has a hydroxyl ion bound to the metal ion. The proton generated is released to bulk water or to external buffer at the rate of the catalysis. This is in most cases the rate-limiting factor. To achieve this, the most active form of the enzyme (isoenzyme II of the α -form) has a histidyl residue in the active site that functions as a local temporary buffer molecule. This histidyl residue is not hydrogen

bonded to other residues and is free to rotate and release the proton generated. This His is not immediately close to the zinc-bound water molecule, but a few water molecules act as bridges for the proton transfer.

The size of the active site pocket does not limit the rate of diffusion of substrate and product molecules. To perform catalysis rapidly, an enzyme should neither bind the substrate nor the products strongly nor require conformational changes. This is the secret of the catalytic rate of carbonic anhydrase. The enzyme forces the substrate to bind in an unexpected way. The active site requires that ligands to the zinc ion are protonated due to the placement of an obligate hydrogen acceptor close to the metal (Figure 8.4). This hydrogen acceptor has been called "the gate keeper" and is the OH of Thr199, where the hydrogen is already occupied in a strong hydrogen bond to Glu 106. This prevents the substrate, bicarbonate, from binding with a negatively charged oxygen to the zinc ion. Instead, the bicarbonate OH group binds to the zinc ion. The bond between the bicarbonate OH group and carbon atom is broken or formed without any rearrangements of the substrate, product or movement of the substrate proton. In addition, bicarbonate is converted to carbon dioxide without any conformational change of the enzyme.

In the formation of bicarbonate, the hydroxyl ion makes a nucleophilic attack on the carbon dioxide. The negative charge is transferred to the carbon dioxide oxygen atom closest to the metal. This oxygen probably becomes a distant ligand to the metal ion through a slight reorientation. Since the bicarbonate cannot bind strongly with a neutral, protonated oxygen at the zinc ion it can rapidly be replaced by a water molecule. The binding of the water molecule and its deprotonization to hydroxyl ion then leads to the dissociation of the product, bicarbonate.

The β -and γ -forms of the enzyme seem to function by the same mechanism even though the structure of the enzyme and the residues in the active site are entirely different.

8.1.1 Transition-State Stabilization

Aromatic sulfonamides are strong inhibitors of all carbonic anhydrases (Figure 8.4). They belong to a classical group of inhibitors that are transition-state analogues in the enzyme reaction. Enzymes generally catalyze reactions by reducing the energetic barriers along the reaction pathway. A transition state of a chemical reaction is a high-energy state that the reactants need to go through to yield the product. Many enzymes catalyze their reactions by stabilizing the transition state and thereby reducing the high-energy barrier. In the reaction catalyzed by carbonic anhydrase, the zinc-bound negatively charged hydroxyl ion reacts with carbon dioxide to produce bicarbonate. The sulfonamide group is an analog of the transition state. The negatively charged and protonated NH group binds to the metal ion and donates a hydrogen bond to the OH of Thr199 while the two oxygen atoms of the sulfonamide group bind in a manner similar to carbon dioxide or the two non-protonated oxygens of bicarbonate. Sulfonic acids on the other hand cannot inhibit the enzyme due to their lack of a proton on the SO₃⁻ moiety.

The enzyme catalysis of carbonic anhydrase becomes very fast due to the simplicity of the reaction and the arrangement of the zinc ion, Thr 199 and Glu106. In mutants where the Thr199 OH has been replaced bicarbonate binds more strongly and two of its oxygens become ligands to the zinc ion. This significantly reduces the catalytic efficiency. Likewise if Glu106 is replaced the efficiency of the enzyme goes down.

Even though α - β -and γ -carbonic anhydrases have very different structures and the residues in the active sites differ, a zinc ion is always a central feature. The ligands to the zinc ion differ in the β -carbonic anhydrases from the other two forms, but in all three forms of the enzyme sulfonamides are strong inhibitors suggesting that the transition states of the catalytic mechanisms are very similar. Thus, one could say that the existence of carbonic anhydrases is an example of convergent evolution — the wheel has been invented five times.

8.2 Ribonucleotide Reductase — Highly Regulated Enzyme

Ribonucleotide reductase (RNR) is the enzyme that replaces the 2'OH of ribonucleotides with a hydrogen atom to form deoxyribonucleotides (Figure 8.5). This is how the building blocks for the synthesis of DNA are produced. The evolution of this enzyme was a crucial step in the transition from the rRNA world to the DNA world. Due to the central role of the enzyme it is an attractive target for cancer and viral therapy. A cysteine residue in the protein, which is converted to a thiyl radical, is a central component in the catalysis.

There are three classes of RNRs: I, II and III (Figure 8.6). They differ in the way they interact with oxygen and the way they generate stable or transient radicals that in turn generate the thiyl radical (Table 8.1). For the multi subunit class I and III enzymes, the thiyl radical is generated by an activating subunit that is not involved in catalysis, whereas in class II it is generated directly on the catalytic subunit by the cleavage of adenosylcobalamin. The required electrons are provided from different sources. Class I requires oxygen for the generation of the radical, class II is indifferent to oxygen and for class III the development of the glycine radical is inhibited by oxygen (Table 8.1). The different classes may relate to the appearance of oxygen during evolution.

Ribonucleotide reductase can reduce all four NTPs or NDPs to dNTP or dNDP. A single enzyme in each organism is able to produce a balanced supply of the four deoxynucleotides. The activity is regulated in such a manner that if there is a shortage of one deoxynucleotide this is the one that will be produced. The control is due to a regulatory or specificity site where the binding of rNTPs or dNTPs regulates the substrate specificity (Table 8.2). The structural background for the activity control is gradually emerging.

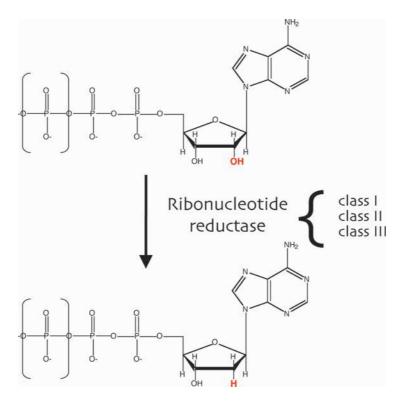


Fig. 8.5 ■ The reaction catalyzed by ribonucleotide reductase (RNR). The 2' hydroxyl group of the ribose of ATP is replaced by a hydrogen atom to become dATP. Other nucleotides will undergo the same change in a controlled fashion. (Illustration kindly provided by Derek Logan.)

8.2.1 Structures of Ribonucleotide Reductase

The structures of the three forms are closely related and the secondary structural elements involved in the allosteric regulation are the same. Thus, the different forms of the enzyme must have been generated through a divergent evolution from the same progenitor enzyme (Figure 8.7). The core of the catalytic subunit is a 10-stranded α/β -barrel for all forms of the enzyme. Two parallel five-stranded sheets joined in an antiparallel way form the barrel.

8.2.1.1 Sites for generation of the radical

The generation of the thiyl radical used in the catalytic mechanism is quite different in the three forms of the enzyme, and this is dependent on the access to oxygen. In the class I enzymes where there are two subunits which handle different parts of the enzymatic process, the radical is generated at a tyrosyl residue close to a Fe-O-Fe complex buried in the R2 subunit. The two subunits do not seem to interact more than transiently. When

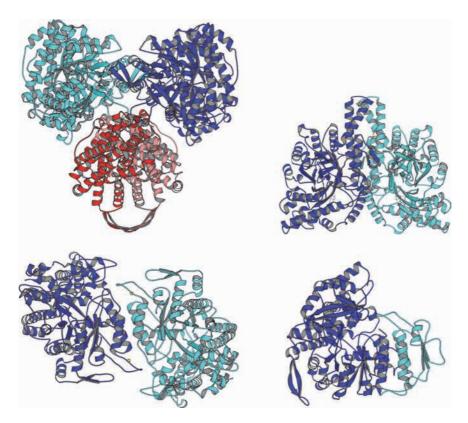


Fig. 8.6 • The three forms of RNR. *Top left*: The Tetrameric class I enzyme from *E. coli*. This is a symmetric complex that has not been observed (based on PDB: 1RLR and 1RIB). *Top right*: A dimer of the catalytic α subunits of the tetrameric class III enzyme from phage T4 (PDB: 1HK8). The structure of the β subunit of the class III RNR is not yet known. *Bottom left*: The dimeric class II enzyme from *Thermatoga maritima* (PDB: 1XJE). *Bottom right*: The monomeric class II enzyme from *Lactobacillus leichmannii* (PDB: 1L1L). The catalytic subunits are blue and turquoise, and the R2 subunit of the class I enzyme is red.

TABLE 8.1 Properties of the Different Classes of Ribonucleotide Reductase

| Property | Class I | Class II | Class III |
|--------------------|----------------------------------|-----------------------------|--|
| Oligomeric state | R1 ₂ *R2 ₂ | α or α_2 | $\alpha_2\beta_2$ |
| Oxygen requirement | Needs oxygen | Indifferent to oxygen | Damaged by oxygen |
| Primary radical | Tyrosine | Adenosylcobalamin | Glycine |
| Radical generator | Fe-O-Fe center | Adenosylcobalamin | [4Fe-4S] center, S-adenosyl methionine, reduced flavodoxin |
| Electron donor | Thioredoxin or glutaredoxin | Thioredoxin or glutaredoxin | Formate |

| | Preferred Substrate | | |
|------------------|---------------------|----------|-----------|
| Specificity Site | Class 1 | Class II | Class III |
| dTTP | GDP | GDP | GTP |
| dGTP | ADP | ADP | ATP |
| dATP | CDP/UDP | CDP/UDP | CTP |
| ATP | CDP/UDP | CDP/UDP | CTP |

TABLE 8.2. The Allosteric Regulation in RNRs

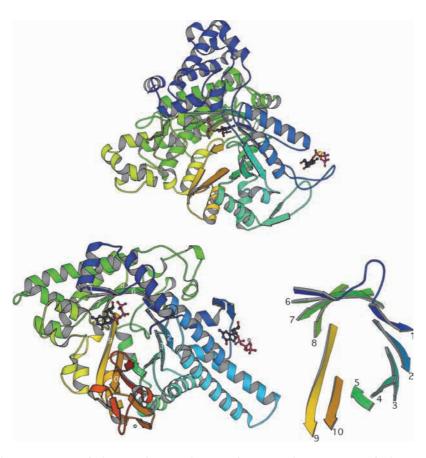


Fig. 8.7 ■ The structure of the catalytic subunit of RNR. The enzyme of classes I and II are represented by the class II enzyme from Thermatoga maritima (top, PDB: 1XJE), while class III (phage T4) is the bottom figure (PDB: 1HK8). From the structures it is evident that the three forms of RNR are evolutionarily related. The strands of the 10-stranded barrel are numbered in the bottom right figure, which shows only the strands of the barrel. The active site is shown with the binding of a nucleotide in the β -barrel in the center, while the specificity site is shown on the right-hand side.

they do the radical has to be transferred, probably as a concerted electron-proton transfer, from the R2 subunit to the cysteine residue in the active site, which is more than 30 Å away in the R1 subunit. The radical in the other classes of RNR does not have to be transferred over long distances. The cobalamin in class II enzymes most likely interacts directly with the cysteine residue to generate the thiyl radical. In the class III enzymes, the glycyl radical is also near the active site.

8.2.1.2 Regulation of enzyme activity

As stated above, RNRs have one, sometimes two, sites that regulate the activity to provide balanced amounts of deoxyribonucleotides for the cell. The overall activity site (Figure 8.8) is not always present. When ATP is bound the activity is stimulated, but when dATP is bound the activity is inhibited. In the inhibited state the enzyme can aggregate into non-productive oligomeric structures. To bind dATP is more deeply into the site than ATP due to the lack of the 2′-OH.

The substrate specificity or effector site is far from the active site within the class I subunit. However, in the dimer it binds across the subunit interface at the active site (Figure 8.8). This arrangement would seem impossible for the monomeric class II version of the enzyme, but a small, inserted domain mimics the essential parts of the substrate specificity site of the missing subunit (Figure 8.6). Conclusive experiments, with nucleotides bound to both sites, have been performed on a dimeric class II enzyme (Figure 8.9).

The sugars and the phosphates of effector and substrate nucleotides are bound in the same way and with the same interactions to the protein in the different complexes. This enhances the possibility for the nucleotide at the effector site to affect the nucleotide

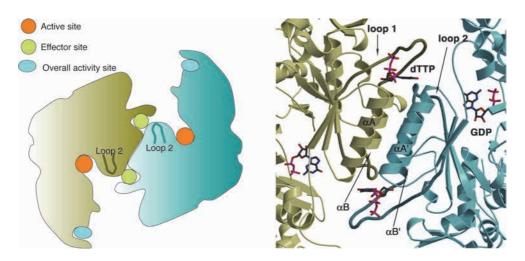


Fig. 8.8. ■ *Left*: The organization of the dimer of the catalytic subunit (R1) of class I RNR. The spatial relationship between the active site and the effector site at the subunit boundary is shown. The overall activity site is present only in some class I and class III enzymes. *Right*: Details of the binding site for effector (dTTP) and substrate (GDP). It should be noted that loop 2 is located between the nucleotide in the substrate specificity site and the one in the catalytic active site. Furthermore, the bases in both sites are oriented towards loop 2. Loop 2 has a major role in the regulation of the substrate selected. The arrangement of these sites in RNRs of other classes is similar. (Illustration is kindly provided by Derek Logan.)

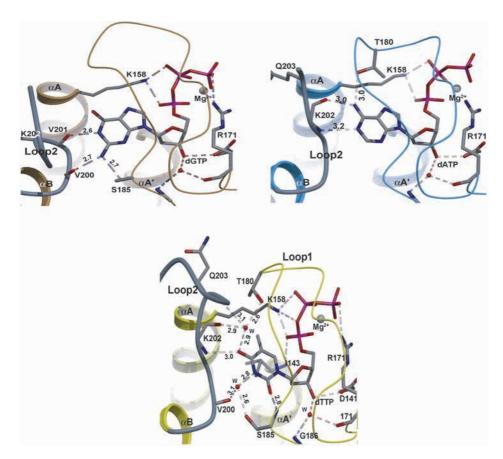


Fig. 8.9 ■ The binding of dNTPs to the effector site of RNR class II (dGTP, dATP and dTTP). The deoxyribose and the phosphates bind in the same manner and fix the nucleotide in a specific location. This has the effect that the nucleotide bases interact very differently with loops 1 and 2, which adopt remarkably different conformations. These interactions are primarily with main chain atoms of loop 2. In this way, the effectors transmit structural signals to the active site, thereby directing the affinity for the nucleotide that should bind and be reduced (PDB: 1XJJ, 1XJF and 1XJM. (Reprinted with permission from Larsson et al. (2004) Structural mechanism of allosteric substrate specificity regulation in a ribonucleotide reductase. Nat Struct Mol Biol 11, 1142-1149. Copyright (2004) Nature.)

binding to the active site. Loop 1 interacts solely with the nucleotide bound at the effector site. However, the loop between αB and βC (Loop 2) is a crucial element that affects the substrate-binding site in relation to the nucleotide bound at the effector site. Loop 2 (residues 199–210 in the class II enzyme from T. maritima) lies between the effector site and the active site across the subunit interface. The loop is very flexible and nucleotides bound to the effector site bind to different main chain groups and lead to conformational changes (Figure 8.9). These conformational changes lead to the nucleotide preferences in the active site (Figure 8.10). In all the complexes, the substrate is clamped by Ala210.

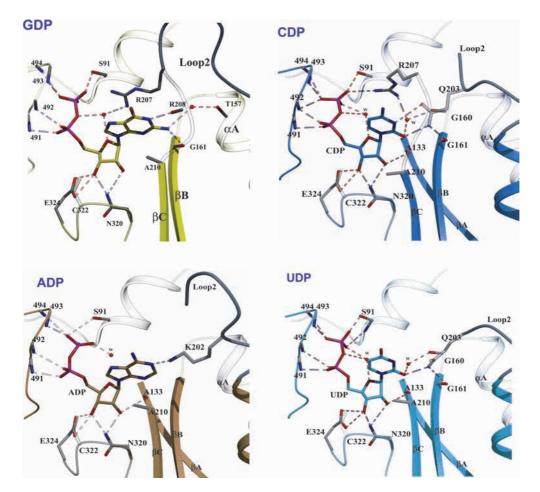


Fig. 8.10 ■ The selection of substrates for the active site of the class II form of the RNR. The ribose and the phosphates are bound in the same way for all four nucleotides. The selection depends on the conformation of loop 2 (residues 199–210), which in turn depends on the nucleotide bound at the effector or substrate specificity site. Interestingly, the recognition of GDP is entirely through main-chain atoms. In this complex, the substrate-proximal side of loop 2 is completely ordered. In the other complexes, different residues from the effector-proximal side of loop 2 protrude through to the active site and the side of loop 2 near the substrate is disordered (PDB: 1XJE, 1XJN, 1XJK and 1XJG. (Reprinted with permission from Larsson *et al.* (2004) Structural mechanism of allosteric substrate specificity regulation in a ribonucleotide reductase. *Nat Struct Mol Biol* 11, 1142–1149. Copyright (2004) Nature.)

The guanidino group of Arg207 sometimes forms a stacking interaction with the bases and a charge interaction with the phosphates. Two key residues are Lys202 and Gln203 that make hydrogen bonds to the substrate bases in different ways for the different substrates. A comparison of the conformations of loop 2 and these key residues is shown in Figure 8.11.

Fig. 8.11 ■ A comparison of the conformation of loop 2 and the locations of the key residues Lys202 and Gln203 between the effector site and the active site for three complexes of RNR class II. (PDB: 1XJN and 1XJK, colors as in Figure 8.10). (Reprinted with permission from Larsson *et al.* (2004) Structural mechanism of allosteric substrate specificity regulation in a ribonucleotide reductase. *Nat Struct Mol Biol* **11**, 1142–1149. Copyright (2004) Nature.)

8.2.1.3 Active site and catalysis

The phosphate and ribose moieties of the four substrates of the class II RNR are bound in the same way in the active site. The thiyl radical carried by Cys439 in this enzyme is close to the 3'-carbon of the ribose (Figure 8.12). On the opposite side of the ribose, two cysteinyl residues (Cys225 and Cys462) are close enough to each other to form a disulfide bond. The two hydroxyl groups of the ribose are hydrogen bonded to Asn437 and the 3'-hydroxyl is hydrogen bonded to Glu441. The thiyl radical transfers the radical to the ribose by abstracting the hydrogen on the 3'-carbon. This leads to the loss of the 2'-hydroxyl to become a water molecule where the missing proton comes from one of the cysteines. Subsequently, the ribose will regain its hydrogen atoms from the cysteines that now form a disulfide bond and the radical is regained. The disulfide bond needs to be reduced for the cysteines to be able to participate in the next catalytic round.

8.3 Motor Proteins and Molecular Switches

The energy currency of the cell is ATP. There is an enormous need for ATP in an organism. A full-grown person utilizes between 50 kg and 75 kg of ATP each day. The ATP is regenerated using chemical gradients across membranes. Such gradients are normally due to different pH or sodium concentrations on each side of a membrane (see Figure 8.15 and Chapter 13). The gradient generates a motion, kinetic energy, which is used to generate

Figure 8.12 ■ The catalytic mechanism for ribonucleotide reductase (after Nordlund–Reichard, 2006). The free radical is shown in green, the ribose OH is shown in red and the atoms and bonds that are different between subsequent steps are highlighted in blue. The numbering above relates to RNR class I. Residues 437, 439 and 441 in this figure correspond to 320, 322 and 324 in RNR class II in Figure 8.10.

chemical energy in the form of ATP. The energy from the ATP is used through hydrolysis to drive a range of molecular processes. Numerous ATPases hydrolyze the ATP and undergo conformational changes needed for different types of work to become motor proteins.

The GTPases are another class of enzymes also using a nucleotide triphosphate, GTP, for their functions. The GTPases or G-proteins, generally function as molecular switches with two main states, ON and OFF, signaling that their duty will be done or that it is done. Some of the GTPases may also function as motor proteins. A characteristic difference between a motor protein and a molecular switch is at what stage the nucleotide is hydrolyzed. A classical motor protein hydrolyses its ATP molecule before the conformational change and thus before the work is done. However, a molecular switch in the ON state induces a process. When this process is completed, the nucleotide is hydrolyzed, leading to a conformational change and a loss of affinity of the molecular switch for the receptor. The hydrolysis of the triphosphates is essentially irreversible because of the repulsion of the products. To regenerate ATP from ADP and P_i requires energy as is exemplified by ATP synthase.

8.3.1 P-Loop-containing Nucleotide Triphosphate Hydrolases

The NTPases all have a central, mostly parallel β -sheet with helices on both sides that connect the strands, the Rossmann fold. Many NTPases belong to a superfamily of proteins that have a loop (the P-loop) with a conserved sequence GXXXXGKT/S, also called the Walker A motif. It interacts with the phosphates of the nucleotides. In particular, the main chain nitrogens of several residues of the P-loop interact with the β -phosphate (see Figure 8.13).

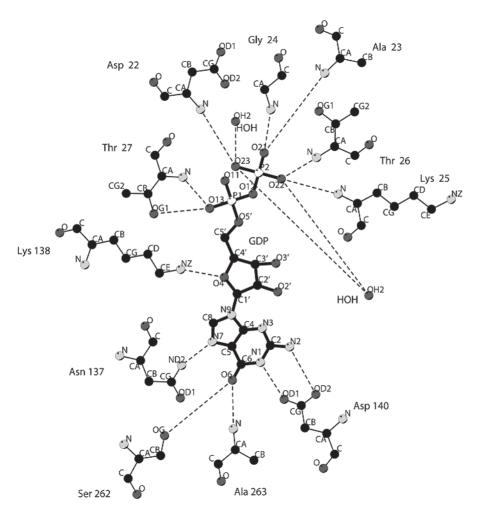


Fig. 8.13 • The interactions between GDP and *T. thermophilus* EF-G. Gly24, Lys25 and Thr26 are some of the conserved residues in the P-loop binding to the α -and β -phosphates. Asn137, Lys138 and Asp140 are part of the NKXD or G4 motif that binds to the guanine. Ser262 and Ala263 of the G5 motif also interact with the guanine. Switches I and II (the G2 and G3 motifs, respectively) interact with the magnesium and γ -phosphate. (Reprinted with permission from Al-Karadaghis S, Ævarsson A, Garber M, *et al.* (1996) The structure of elongation factor G in complex with GDP: conformational flexibility and nucleotide exchange. *Structure* 4: 555–565. Copyright: Elsevier).

The NTPases can also have a Walker B motif with one or two acidic residues preceded by four hydrophobic residues. In many P-loop proteins these motifs are found in loops connecting neighbor β -strands with α -helices. However, the topology of the sheet and the presence of further elements differ between different families of this group. Some examples of these enzymes are described below. More details can be found in the sections indicated in Table 8.3.

TABLE 8.3 Families of P-loop NTPases and Examples of Proteins in These Families. The Strand order is Focusing on the core Conserved Elements and the Strand Preceding the P-loop is Given Number 1

| Domain Family | Strand Order | Protein Examples | Oligomeric State | Water Activation | Sections |
|------------------|--------------------------------|--|--------------------------------|---------------------|----------|
| RecA-like | 23415 | Bact. RecA, recombinases | Right-handed helical filaments | | 8.3, 9.4 |
| | 32451 | F_1 -ATP synthase, subunits α, β | Quasi-hexamer | Glu | 8.3 |
| | | DNA helicases | Monomers, two RecA domains | | 9.2 |
| | | DNA helicases (DnaB) | Hexamer, one RecA domain | | 9.2 |
| | | ABC transporters | Dimer | | 13 |
| AAA+ | 23415 | Origin of replication (DnaA) | Quasi-hexamer | | 9.2 |
| | | DNA helicases | Hexamer | | 9.2 |
| | | Helicase loader (DnaC) | Hexamer, not symmetric | | 9.2 |
| | | Clamp loader | Pentameric ring | | 9.2 |
| | | Hsp100, ClpA-C, B-C, Lon | Hexamer, two AAA+ domains | | 12 |
| AAA | | Clp-A-N and B-N, FtsH, Hsp104, Hsp78 | Hexamer | | 12 |
| Motor proteins | 2314, strand 2 is antiparallel | Myosin, kinesin heavy chains | | | 15, 18 |
| Kinases | 23145 | | Monomer | | 14 |
| G-domains | 231456, strand 2 | Ras (p21) | Monomer | Gln | 8.3, 14 |
| | is antiparallel | Trimeric G-proteins, α subunit | Heterotrimer | Gln | 14 |
| | | Translation factors (EF-Tu, EF-G) | Monomer | His | 11 |

8.3.1.1 RecA and RecA-like proteins

RecA is a bacterial protein of about 38kDa involved in recombination of dsDNA molecules of similar sequence. RecA is composed of a core domain with the ATP-binding site with smaller N- and C-terminal domains. Rad51 is the corresponding protein in eukaryotes and archaebacteria (Figure 8.14). It is the prototype of recombinase proteins in most living organisms. RecA polymerizes in a helical manner on ssDNA and engages in strand exchange (see Section 9.4). The ATP-binding site is usually between subunits with residues from both subunits participating in catalysis. In particular, an arginine residue, called the Arg finger, which is involved in stabilizing the γ -phosphate in the transition state of catalysis is often part of a neighbor subunit.

RecA-like domains are found in a range of enzymes that use ATP for different kinds of mechanical work. The central β -sheet of RecA has the order of the parallel β -strands: 32451678, where strand 7 is antiparallel. If one considers strands 3, 7 and 8 as additions to a more basic motif the strand order becomes 23415. This is the order common for a wide range of NTPases. Due to similar structural features common to smaller groups of enzymes several subfamilies can be identified. The enzymes perform work through conformational changes between the ATP and ADP states.

The RecA-like proteins often form hexameric rings (Table 8.3), but sometimes they are monomeric with two RecA-like domains. The ATP is bound between subunits or in the case of monomers between domains. In RecA proteins, the strands associated with the Walker A and B motifs are separated one additional strand (number 4) containing what has been called sensor-1 in the proximity of the bound ATP.

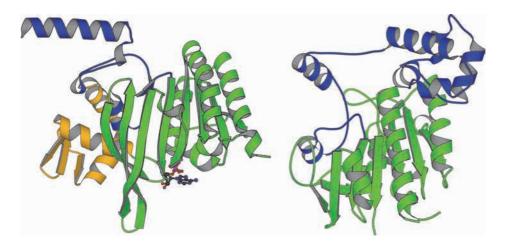


Fig. 8.14 ■ *Left*: The structure of RecA from *E. coli*. The core domain shown in green, is the prototype of the RecA fold found in many proteins. A bound ADP is also shown. The N-terminal domain is in blue and the C-terminal domain in orange (PDB: 1REA). *Right*: The human Rad51 protein has the same topology of the central sheet but the N-terminal domain is different (PDB: 1SZP).

8.3.1.2 AAA and AAA + proteins

One large group of RecA-like ATPases is called the AAA+ proteins (ATPases associated with various cellular activities). They perform different type of mechanical work and frequently form hexameric rings. The AAA+ domain with the ATP-binding site has 200–250 amino acid residues. A defining feature of AAA+ enzymes is the placement of the arginine finger on the C-terminal side of a helix that precedes strand 5. The arginine is directed to the ATP of a neighbor subunit. In this way, strand 4 forms a link between two neighboring ATP sites and can communicate the state of the nucleotide of one subunit to the next subunit, important for the conformational changes involved in the helicase, protein disassembly or other physical work performed by the AAA+ molecules (Table 8.3).

The classical AAA proteins are a subgroup of the AAA+ family of ATPases. A characteristic structural feature is a small helix between strand 2 and helix 2 and a GNR motif associated with the Arg finger. Members of this family are FtsH and the N-terminal ATPase domains of ClpA and B. CplA has double AAA+ domains where the N-terminal domains belong to the AAA family but the C-terminal domains have the characteristic features of AAA+.

The enzyme ATP synthase has three α and three β subunits all having AAA+ domains (Table 8.3).

8.3.2 ATP Synthase

ATP is produced in mitochondria, chloroplasts and bacteria by ATP synthase (F_1F_0 -ATPase). ATP synthase is a ubiquitous motor protein conserved from bacteria to humans. The synthesis of ATP from ADP and phosphate, catalyzed by ATP synthases, is the most abundant physiological reaction in almost all cells. The systems involved in generating the ion gradient and ATP synthase are shown in Figure 8.15. ATP synthase is a member of a family of rotatory ATPases, which also include the vacuolar H^+ -ATPases (V-ATPases) and A_1A_0 -ATPases (A-ATPases) that can function both as ATP synthases and ion pumps. Our main focus will be on the F_1F_0 -ATPase. All forms of rotatory ATPases are composed of a membrane bound unit (e.g. F_0) and a soluble unit (e.g. F_1). Either unit can drive the other depending on the direction of the ion flow.

8.3.2.1 Chemiosmotic theory. All you need is — energy

The cell is powered not by chemical reactions, but by a kind of electricity, specifically by a difference in the concentration of protons, H⁺, across a membrane. Because protons have a positive charge, the concentration difference produces an electrical potential difference between the two sides of the membrane of about 150 millivolts. This may not sound like much, but since it operates over only about 50 nm, the field strength

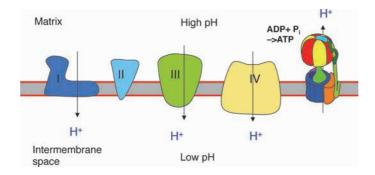


Fig. 8.15 ■ In the respiratory chain, protons are pumped across the membrane by complexes I (NADH dehydrogenase), III (cytochrome bc_1 complex) and IV (cytochrome c oxidase) to generate a proton gradient across the membrane. This gradient drives the synthesis of ATP by the ATP synthase. Complex II (succinate dehydrogenase) does not pump protons. The ATP synthase (on the *right*) uses the proton gradient generated by the other enzymes to synthesize ATP.

over the tiny distance is enormous, around 30 million volts/meter. That is equivalent to a bolt of lightning. This electrical driving force is called the proton-motive force (see below).

Essentially, all cells are powered by this force field that is as universal to life on Earth as the genetic code. This tremendous electrical potential can be tapped directly, to drive, for instance, the motion of bacterial flagella or harnessed to produce the energy-rich fuel ATP.

However, the way in which this force field is generated and tapped is extremely complex. The enzyme that makes ATP is a rotating motor powered by the inward flow of protons. Another protein that helps to generate the membrane potential, NADH dehydrogenase, is like a steam engine, with a moving piston for pumping out protons. These amazing nanoscopic machines must be the product of a long natural selection. The process of ATP synthesis is called oxidative phosphorylation:

$$ADP + P_i \Leftrightarrow ATP + H_2O$$
 $\Delta_r G^{\oplus} = +31 \text{ kJmol}^{-1}.$

The synthesis of ATP in the cells is driven to produce ATP concentrations far above the equilibrium concentration and is used by numerous cellular processes that require ATP for their function. ATP is synthesized directly from ADP and inorganic phosphate, P_i , in bacterial cells or mitochondria. For many years it was a mystery how this might happen. In 1961, Peter Mitchell (Nobel laureate for chemistry, 1978) proposed the following essential features:

- The energy transfer intermediate, linking oxidations and ATP synthesis, is a transmembrane ion gradient.
- The ions involved in the gradient are protons (or H₃O⁺).

These two concepts form the essential postulates of the chemiosmotic theory. The term "chemiosmosis" signifies the link between chemical reactions and energy stored in a transmembrane gradient ("osmotic energy"). Thus, according to Mitchell, the energy released during the transport of electrons along the carrier chain is conserved in the form of an H⁺ gradient and an electrical gradient that then drives the oxidative phosphorylation. As the electrons flow down the electron transport chain, H⁺ is expelled from the inside membrane of the mitochondrion to the intermembrane space (Figure 8.15). This results in a rise in pH on the inside and a fall in pH on the outside of the inner membrane — a pH gradient is generated and maintained. The electric potential also rises across the membrane because more positive H⁺ ions are on the outside than on the inside. The protons on the outside have a thermodynamic tendency to flow back in, so as to equalize the pH on the two sides of the membrane. Another way to put it is that Gibbs energy must be expended to maintain the proton gradient. When protons flow back into the inner of the mitochondrion their energy is dissipated and some of it is used to drive the synthesis of ATP. The force can also be generated by a sodium gradient across the membrane in some microorganisms.

The proton motive force is generated across the inner membrane of mitochondria in humans and animals, in the inner membrane of chloroplasts in plants, and over the plasma membrane of aerobic bacteria. The chemiosmotic mechanism thus depends on a pH gradient being established across the membrane by a series of electron transfer reactions. The ATP synthesis is driven by the concurrent transport of protons across the membrane. A simplified overall reaction can be written as the sum of two reactions:

$$\begin{split} & \text{ADP} + \text{P}_{\text{i}} \Leftrightarrow \text{ATP} + \text{H}_{2}\text{O} \\ & \frac{x\text{H}_{\text{out}}^{+} \Leftrightarrow x\text{H}_{\text{in}}^{+}}{\text{ADP} + \text{P}_{\text{i}} + x\text{H}_{\text{out}}^{+} \Leftrightarrow \text{ATP} + x\text{H}_{\text{in}}^{+} + \text{H}_{2}\text{O}}. \end{split}$$

The energy stored in a transmembrane H^+ gradient then comes from two contributions. First, the difference in activity of H^+ ions results in a difference in the chemical potential across the cell membrane:

 $\Delta G_m = G_{m,in} - G_{m,out} = RTln(a_H^+,_{in}/a_H^+,_{out})$ that comes from the difference in entropy of mixing in the two compartments.

Second, there is a membrane electric potential difference $\Delta \phi = \phi_{in} - \phi_{out}$ that arises from differences in electrostatic interactions on each side of the membrane, the outside being positive and the inside negative. The charge difference across a membrane per mole of H⁺ ions is $N_A e = F$, Faraday's constant and ΔG for this process is equal to $F\Delta \phi$.

Therefore, for each proton transported from the inside to the outside, the total Gibbs energy stored is (i.e. the Gibbs energy available for phosphorylation according to the chemiosmotic theory):

$$\Delta G_{trans} = F\Delta \phi - (RTln10)\Delta pH$$

where we have replaced activities with molar concentrations and introduced $pH = -\log[H^+]$ and substituted $\Delta pH = pH_{in} - pH_{out} = -\log[H^+]_{in} + \log[H^+]_{out}$.

In the mitochondrion $\Delta pH \approx -1.4$, corresponding to a ΔG of 8 kJmol⁻¹ at 25°C, and $\Delta \phi \approx 0.14$ V, corresponding to a ΔG of 13.5 kJmol⁻¹ at 25°C, so $\Delta G_{trans} = 21.5$ kJmol⁻¹ at 25°C. Since 31 kJmol⁻¹ is needed for phosphorylation, at least 2 mol H⁺ (and probably much more since we disregard the passive proton leakage through the membrane) must flow through the membrane for the phosphorylation of 1 mol ADP. The effective electrical capacitance and buffer capacity of the mitochondrion determine the distribution of free energy between $\Delta \phi$ and ΔpH .

8.3.2.2 Structure

ATP synthase has been observed by EM as distinct lollipop-like structures associated with mitochondrial membranes (Figure 8.16).

ATP synthase has two separate units, F_0 (the letter "o", not the digit "zero") and F_1 . The former is transversing the membrane while the F_1 unit is located on the cytoplasmic side of the membrane. The subunits either belong to the rotor or stator part of the motor (Table 8.4). During its function, the proton gradient across the membrane forces the rotor of the ATP synthase to rotate while the stator parts remain fixed (Figure 8.17).

The F_1 unit consists of three copies each of two subunits called α and β , and one copy each of subunits γ , δ and ϵ . The active sites for ATP synthesis are found on the three β subunits. They form a hexameric aggregate alternating with the three α subunits that are similar to the β subunits. The γ subunit is threaded through this hexamer. It is a left-handed α -helical coiled-coil. At its broader base, the δ and ϵ subunits are attached (Figure 8.17). Due to its asymmetry, the γ subunit has different contacts with the three β subunits and forces them to adopt different conformations (Figures 8.18 and 8.19).

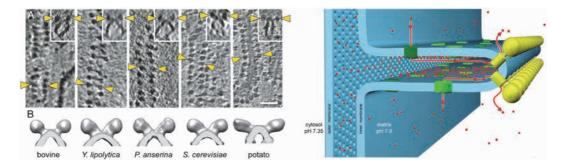


Fig. 8.16 ■ *Left.* Under the electron microscope, lollipop structures are observed emerging at the tips of cristae of the inner membranes from mitochondria (see arrows). These are ATP synthase. *Right.* The cristae membranes are organized with ATP synthase (yellow) at the tips and proton pumps of electron transport complexes, primarily complex I (green), in adjacent regions. Protons (red) are pumped into the cristae to be used by the ATP synthase. (Reproduced with permission from Davies *et al.* (2011) Macromolecular organization of ATP synthase and complex I in whole mitochondria. *PNAS* **108**: 14121–14126.)

| 2011) | | | | | | |
|-------------------------|---------------------------------------|---------------------------------------|---------------------------------------|-----------------|--|--|
| V-ATPase | A ₁ Ao-ATPase ^b | F ₁ F _o -ATPase | F ₁ F _o -ATPase | Comment | | |
| Eukaryotic ^a | Archaea | Mitochondria | Bacteria | | | |
| 3 A | 3 A | 3 β | 3 β | Stator | | |
| 3 B | 3 B | 3 α | 3 α | Stator | | |
| D | D | γ | γ | Rotor | | |
| 3 E | 2 E | OSCP | δ | Stator | | |
| F | F | _ | _ | Rotor | | |
| | | δ | ε | Rotor | | |
| 3 G | 2 G | b | 2 b | Stator | | |
| a | I | a | a | Stator | | |
| c | С | 10 c | 10–15 с | Rotor | | |
| d | С | _ | _ | Rotor | | |
| C, H, e | | ε, d, F ₆ , A6L | | Unique subunits | | |

TABLE 8.4 Common Subunits of F-, A- and V-ATP Synthase (From Muench et al., 2011)

^bThe archaeal A₁A₀-ATPase can function as an ATP synthase or an ion pump.

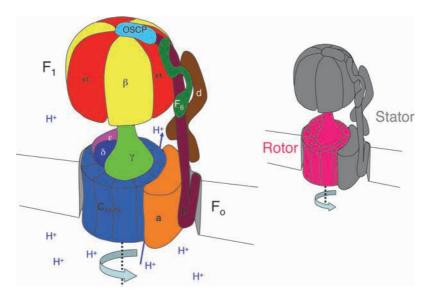


Fig. 8.17 • Organization of mitochondrial ATP synthase. The proton gradient across the membrane drives part of the complex (the rotor) to rotate to transport protons across the membrane at the same time as ATP is generated. The F_0 part is associated with the membrane, whereas the F_1 part emerges from the membrane. *Left*: The ring of c subunits (blue) of F_0 rotate in the membrane and with it the γ (light green), δ (dark blue) and ε (deep purple) subunits of F_1 . The a (orange), b (dark red), d (brown), F_0 (green) and OSCP (blue) of F_0 and the three α (olive green) and three β (light green) subunits of F_1 form the stator. *Right*: The stator parts are shown in grey whereas the rotor parts are shown in red.

^aThe V-ATPase is a vacuolar ATP driven proton pump.



Fig. 8.18 • Detailed structure of the F_1 part of ATP synthase. The green γ subunit is a stalk inside the $\alpha_3\beta_3$ oligomer with three-fold symmetry. The γ subunit is asymmetric and has different contacts with the three catalytic β subunits (PDB: 1E79).

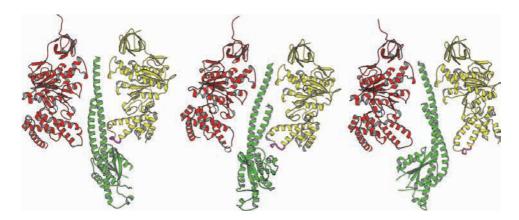


Fig. 8.19 • The asymmetric γ subunit has different contacts with the three α and three β subunits as seen by the three diametrically opposed pairs of α and β subunits. The active sites of the β subunits have three different conformations called β_E (empty), β_{DP} (a bound ADP), and β_{TP} (with a bound ATP). A loop from the C-terminal domain (purple) interacts with the coiled-coil part of the γ subunit and affects the conformation of the active site of the β subunit.

The α and β subunits both have three domains: an N-terminal β -barrel domain close to the top of the molecule (Figure 8.19); a nucleotide binding catalytic domain in the middle and the C-terminal helical domain closest to the membrane. The catalytic site is located at the interface between the α and β subunits. The enzyme is related to the family of enzymes called AAA+ ATPases (Table 8.3). The N-terminal domains form a ring with pseudo-six-fold symmetry. The ring contributes significantly to the stability of the $\alpha_3\beta_3$ complex.

The F_o part differs somewhat in different organisms. In bacteria and chloroplasts it is composed of subunits a, b_2 and c_{10-15} (Table 8.4). In mitochondria F_o has only one copy of b but subunits δ and F_o partly replace the deleted copy of b. Subunits b, d, F_o and OSCP belong to the peripheral or stator, stalk of the motor. Subunit OSCP in mitochondria (δ in bacteria and chloroplasts) attaches to the N-terminal region of one or several α subunits preventing the α and β subunits of F_1 from rotating.

The V-ATPases and the A-ATPases are distinctly different with regard to the stator connections between the membrane part and the soluble part of the enzyme. The dimers of subunit b in bacteria correspond to heterodimers of subunits E and G in V-ATPases and A-ATPases. In A- and V-ATPases, there are two or three respectively peripheral stalks between the F_0 part and the F_1 part.

In the C-terminal domain of the β subunit a loop connecting two helices interacts with the γ subunit (Figure 8.19). This affects the conformation of the β subunits and the state of its active site. A part of this loop has the conserved sequence DELSEED. The first two residues (Asp394 and Glu395) interact with numerous positive charges of the γ subunit (Figure 8.20). These positive charges, some of which are highly conserved, provide a low-energy pathway for the negative charges during rotation of the γ subunit.

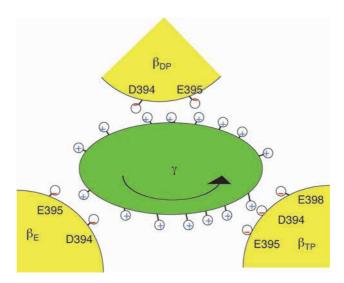


Fig. 8.20 • The coiled-coil part of the γ subunit (green) has a number of positively charged residues that interact with negative charges of the β subunits (yellow).

8.3.2.3 ATP hydrolysis and ATP synthesis

The F₁ part of the ATP synthase can also function as an ATPase. For the catalytic mechanism, Paul Boyer concluded that during ATP hydrolysis the three active subunits must undergo successive conformational changes due to the hydrolysis of ATP. Only a rotatory mechanism could lead to such successive conformational changes. The full enzyme synthesizing ATP due to the proton-motive force must undergo similar conformational changes (Figure 8.21). This "the binding change mechanism" was partly verified by the crystal structure of F_1 , where the asymmetric γ subunit forces different conformations on the β subunits (Figure 8.19).

The definite verification of the rotatory property of the ATPase came with the experiment illustrated in Figure 8.22. The β subunits were bound with His-tags to a Ni-coated cover slip. In the opposite end of F_1 , the γ subunit had a fluorescently labeled actin filament attached. These showed up as fluorescent rods in the microscope. When ATP was added, the actin rods rotated anticlockwise as viewed from the membrane side, making the molecular events of the ATPase visible. When analyzed further the rotation proceeded in a stepwise fashion composed of a 80° step with a short dwell time during which hydrolysis of the ATP occurs and a subsequent 40° step leading to the release of ADP and inorganic phosphate. After 120° of rotation, one ATP molecules was consumed. The β subunits were now jointly at a state identical to that before the rotation, but the individual β subunits had proceeded one step further in the completion of a 360° turn (Figure 8.21). In addition to the rotation during hydrolysis, F_1 alone can synthesize ATP if the γ subunit is forced to rotate in the clockwise direction within the α/β unit.

The ATP synthetase is related to the AAA+ family of enzymes, with a P-loop with the highly conserved sequence GGAGVGKT interacting with nucleotides. Figure 8.23 shows the active site of the β subunit.

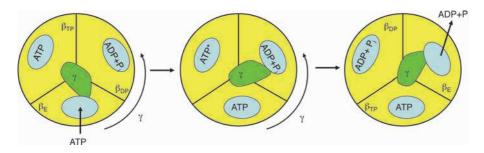


Fig. 8.21 ■ The hydrolytic mechanism of ATP synthase initially proposed by Boyer and subsequently modified according to crystallographic results. The three β subunits are shown in three different conformations during the functional cycle. The β_E subunit is empty, β_{DP} contains an already hydrolyzed ATP molecule and β_{TP} has an ATP molecule bound. When ATP binds to β_{E} , the γ subunit rotates 80° with respect to the β subunits. This leads to an activation of the ATP in the former β_{TP} subunit. When the ADP and inorganic phosphate (P_i) is released, the continued rotation of 40° of the γ subunit leads to the open conformation β_F . ATP synthesis proceeds in the opposite direction, driven by the proton gradient across the membrane.

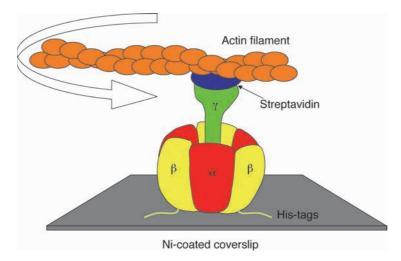


Fig. 8.22 • The proof of the rotatory mechanism was a microscopy experiment where the F_1 unit was attached to a Ni-coated cover slip by extending the N-termini of the β subunits with His-tags, which have a strong affinity to Ni. The rotation of the stalk was observed using a fluorescently labeled actin filament. The actin filament was bound using the protein streptavidin, which binds with high affinity to biotin groups attached to the actin filament and to the γ subunit. By adding ATP, the actin filaments could be seen rotating counterclockwise when seen from the membrane and at a rate depending on the ATP concentration.

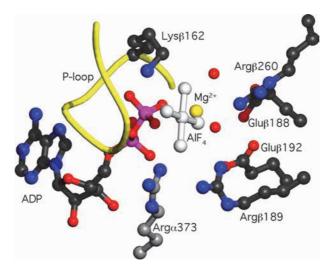


Fig. 8.23 • The active site of the β subunit in the β_{DP} conformation. The ATP is imitated by an ADP molecule and an aluminum fluoride ion bound in the position of the γ -phosphate. The "ATP" interacts with the P-loop K162 and several arginine residues. The α subunit contributes with R373 (grey carbon atoms) which corresponds to the "arginine finger" in RecA-like ATPases. E188 is the residue that activates the water molecule in the hydrolysis of the γ -phosphate.

How can the enzyme synthesize ATP in conditions where hydrolysis would be the natural reaction and how can the enzyme avoid inhibition by the ATP? The affinity for binding ATP or ADP+P_i to the β_{TP} site is about equal, whereas the β_{DP} site very much favors the binding of ADP+P_i. In the hydrolysis reaction, ATP binds to the open β_E site but a large gain in binding free energy is obtained when this site converts to the β_{TP} state. The binding of ATP to a new β_E site drives the ATP in the β_{TP} site to the β_{DP} state. Hydrolysis to ADP+P_i then occurs because the β_{DP} state favors ADP+Pi compared to ATP. This conformational change forces the γ subunit to rotate with respect to the α and β subunits.

ATP synthesis is driven by a clockwise rotation of the γ subunit. The β subunit in a half closed state, β_{HC} , has affinity for ADP+P_i but less for ATP. The rotation of the γ subunit induces the β_{DP} state and, after another rotational step, the β_{TP} state, which leads to ATP synthesis. After yet another rotation of the γ subunit we are back at the β_E state, which is open and has low affinity for ATP. It therefore dissociates. This process is not inhibited by high concentrations of ATP.

A number of key residues in the catalytic mechanisms in the bovine mitochondrial ATP synthase are shown in Figure 8.23. Lys162 is the P-loop lysine conserved in the AAA+ and RecA families. Glu188 interacts with the water that participates in ATP hydrolysis or is a product in the synthesis direction. The only functional residue in the active site from the α subunit is Arg373, which acts as the arginine finger.

8.3.2.4 Motor enzyme

How does the proton motive force generate the rotation that leads to ATP synthesis? The pH gradient over the membrane operates on the F_o part. The γ and ϵ subunits of F_1 bind to the ring of c subunits. The number of c subunits is variable. In the yeast enzyme, there are 10 subunits, but 11 in the Na⁺ ATP synthase from *Ilyobacter tartaricus* (Figure 8.24) and species with up to 15 subunits have been found. Thus, the symmetry of the c-ring frequently deviates from the three-fold symmetry of the α and β subunits. ATP synthases operating at a low ion-motive force tend to have the advantage of a large number of c subunits. Organisms with a constantly high ion-motive force have c-rings with fewer subunits.

The c subunit is normally composed of a helical hairpin spanning the membrane. The residue involved in the binding of the proton is an aspartate halfway through the membrane structure. Figure 8.24 illustrates a situation where an Na⁺ gradient is driving the motor. The sodium ion binds midway on the c subunit and at the subunit interface. A glutamate, a glutamine and a serine are the main ligands to the metal ion. The glutamate corresponds to the proton-binding aspartate.

The a subunit forms the channel for the protons or ions that generate the rotatory work. The structure of this subunit is not known in detail but the channel appears to be composed of two half-channels that are not connected (Figure 8.25). The proton or ion-binding

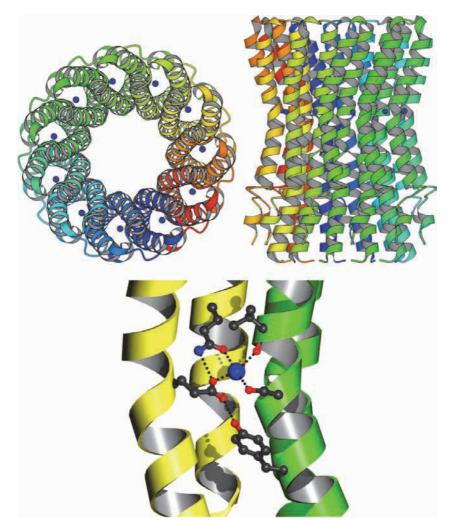


Fig. 8.24 ■ The structure of an ATP synthase from *I. tartaricus* driven by Na⁺. Here, 11 csubunits form the ring spanning the membrane. *Top*: View perpendicular and tangential to the membrane. *Bottom*: A sodium ion is bound between two subunits (yellow and green) (PDB: 1YCE).

site of the c subunits is loaded with a proton or a sodium ion through the inlet channel on the side with the higher concentration of ions. The ring of c subunits are then induced to rotate almost a complete turn to release the proton or ion through to the outlet channel on the low concentration side. During a short passage when the c subunit is close to the a subunit of the stator (Figure 8.25), an arginine of the stator interacts with the glutamate (aspartate) that holds the sodium ion (proton). The ion is thereby released into the outlet channel and a new ion is picked up at the inlet channel. The mechanism can be described as a "push-and-pull" mechanism where the arginine of the a subunit and the proton or sodium ion will compete for interaction with the acidic side chain.

ATP synthesis not only needs a proton or sodium ion gradient but also the membrane potential. In a relaxed mode, the motor can rotate in both directions and perform ion

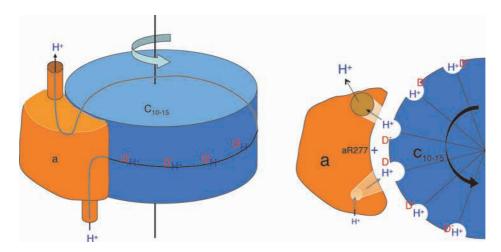


Fig. 8.25 ■ Left: Parts of the F_o unit where the c subunits (blue) are made to rotate in the membrane. The a subunit forms half a channel for the protons (or sodium ions) through which they reach an aspartate on the c subunits (orange) where they bind. To escape, the protons (or sodium ions) force the ring of c subunits to rotate by a ratchet mechanism, almost a complete turn, where they find another half channel through the a subunit. This arrangement is similar to a water mill where different water levels (protons or sodium ions) drives the wheel around to generate work. Right: The "push-and-pull" mechanism of ATP synthase. The aspartate side chain (red D) of the c subunits interacts temporarily with an arginine of the a subunit of the stator. This makes the aspartate lose its proton through the outlet channel (upper). A new proton will bind to the negatively charged aspartate of the next c subunit through the inlet channel (lower). Thus the excess of protons on the outer side of the membrane drives the "wheel" of c subunits and leads to the synthesis of ATP.

exchange between the two sides of the membrane. With a membrane potential and a proton or sodium ion gradient the enzyme switches to unidirectional rotation.

The rotation of the c subunits also involves the γ , δ and ϵ subunits, the rotor part of the motor enzyme. While the γ subunit forms the central or rotatory stalk of the enzyme, the peripheral stalk (or stalks) unites the stator components of the F_o and F₁ parts of the enzyme. The stator comprises subunits a, b, d, F₆ and OSCP of the F₀ unit as well as the α and β subunits of the F_1 unit.

The ring of c subunits rotates in 10 to 15 steps for a full turn while binding and releasing protons or ions. Thus, 3 to 5 protons are used to synthesize one ATP molecule. In a full turn, three ATP molecules are synthesized. The stepwise rotation of the disk of c subunits generates a torque in the γ subunit and the peripheral stalk that is released when the γ subunit passes the β subunit. The torque differs between different species and between the steps of the rotation. The release of the strain is, which is done in two identified steps, 80° and 40° , amount to a third of a complete turn and the production of one ATP molecule.

The structures of more complex rotatory systems are gradually being unraveled, e.g. the flagellar motor that functions as a propeller to drive bacteria forwards in their media.

8.3.3 G-proteins or GTPases

The G-proteins or GTPases are a large family of enzymes and are related to the RecA-like family. The GTP-binding domains or G-domains have the same general fold and bind the nucleotide in the same manner. The G-proteins are molecular switches. They have an ON state (with GTP) and an OFF state (with GDP). The hydrolysis of the GTP to GDP and inorganic phosphate (P_i) is generally induced through interaction with other protein molecules called GTPase-activating proteins (GAPs). To catalyze the release of GDP and to promote the binding of a new GTP molecule many of these proteins interact with a specific G-nucleotide exchange factor (GEF). Examples of G-proteins interacting with their respective GAP and GEF proteins are found in the sections on translation (Chapter 11) or signaling (Chapter 14).

The G-domain normally has 160–200 amino acid residues and belong to the RecA-like family of enzymes (Table 8.3). The G-proteins can be classified as several subfamilies. The simple G-protein Ras used in some signal transduction pathways is monomeric with only one domain, the G-domain (see Chapter 11), while other members are multidomain proteins that contain a G-domain. Another subfamily is the heterotrimeric G-proteins used in other signaling pathways.

8.3.3.1 Structure. The consensus elements. Nucleotide and Mg^{2+} binding

The Ras protein is formed by a mostly parallel sheet with helices on both sides (Figure 8.26). It is similar to the Rossmann fold found in many nucleotide-binding enzymes, but has the strand order 231456, where strand 2 is antiparallel to the others (the Rossmann fold has strand order 321456 with all strands parallel).

The core of the G-domain is well conserved in all G-proteins, but variations in the domain are frequent. Five consensus elements or motifs (G1 to G5) are characteristic of G-domains (Table 8.5). The G1 and G2 are identical to the Walker A and B motifs described in Section 8.3.1.

The five consensus elements primarily connect the C-terminal side of the β -strands to the α -helices and constitute the binding sites for the nucleotide and a magnesium ion (Figure 8.27). Generally, the three first consensus elements, the P-loop, switches I and II, interact with the GTP/GDP phosphates and the magnesium ion, while the two last, G4 and G5 control the selectivity of the G-nucleotides. The magnesium ion is an essential cofactor for GTP hydrolysis.

The first loop (G1) is called the phosphate-binding loop or P-loop (Section 8.3.1). It folds around and makes hydrogen bonds to the α - and β -phosphate moieties of the G-nucleotide. Switch I (G2), or Walker B or effector region, and switch II (G3) primarily interact with the β - and γ -phosphates of the nucleotide through the magnesium ion. The conformations of these loops respond to whether a GTP or a GDP molecule is bound or

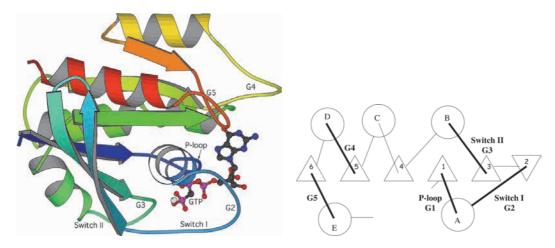


Fig. 8.26 ■ Left. The G-domain — the example chosen is from the Ras protein that is active in many signal transduction pathways. The figure is colored from N-terminal (blue) to C-terminal (red). Right: The organization of the G-domain. The circles represent α-helices and the triangles β-strands. The numbers on the secondary elements indicate the order number of helices and strands, respectively. The five loops with the consensus elements are highlighted. The P-loop, switches I and II are essential for interaction with phosphates and recognition of γ-phosphate. The G4 and G5 loops identify the G-nucleotide.

TABLE 8.5 Consensus Elements of the G-proteins

| Element | Alt. Name | Sequence | Role |
|-----------|-----------|------------|--|
| P-loop | G1 | GXXXXGKT/S | Interactions with α- and β-phosphates |
| Switch I | G2 | XTX | Binding of γ-phosphate and Mg ²⁺ |
| Switch II | G3 | DXXG | Binding of γ-phosphate and indirect Mg ²⁺ binding |
| _ | G4 | N/TKXD | Recognition of the G-nucleotide |
| _ | G5 | T/GC/SAL/K | Binding of the G-nucleotide |

whether the nucleotide-binding site is empty. The effector loop (switch I) is involved in receptor binding and switches conformation drastically between the GDP and GTP states. Switch II relays the status of the bound nucleotide to the conformation of multidomain GTPases.

G-proteins are called molecular switches. When the G-protein with a bound GTP molecule is in the ON state, it can bind to a receptor or effector. This interaction may lead to an interaction with GAP that induces the G-protein to hydrolyze its bound GTP molecule. After the GTP hydrolysis, the conformation of the G-protein changes to its OFF state and dissociates from the effector. Evidently, G-proteins are incomplete enzymes. They have a low intrinsic GTPase activity. Therefore, they need to interact with the appropriate components of the cell to become activated.

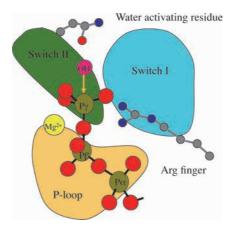


Fig. 8.27 The P-loop contacts the α - and β -phosphates. The switch I and II loops of the G-domain interact with the γ -phosphate and magnesium ion and adopt the ON conformation. A GTPase-activating protein (GAP) induces an active GTP-hydrolyzing conformation in the G-protein. Part of this activation is due to an arginine residue (Arg finger) that interacts with the γ -phosphate and stabilizes the transition state. In this process, the water molecule is moved closer to the γ -phosphate, deprotonated and induced to react with the phosphate in an in-line sn² mechanism.

8.3.3.2 GTP hydrolysis

The mechanism of GTP hydrolysis is thoroughly studied in many G-proteins. The role of the GAP is to induce a conformational change in the protein that makes it an active GTPase and release it from its effector. The GTPase induction is probably similar for all GTPases. A water molecule or more specifically, a hydroxyl ion, suitably placed near the γ -phosphate opposite the β -phosphate, is needed for the GTP hydrolysis. Such water molecules have frequently been seen in GTPase structures. Two requirements for the activation of a GTPase are the stabilization of the transition state and the activation of the water molecule by removing a proton. This deprotonation allows the water molecule to make an associative in-line sn² attack on the γ -phosphate and hydrolyze the GTP to GDP and inorganic phosphate (Figure 8.28). An alternative mechanism is the dissociative hydrolysis, where the dissociation of the γ -phosphate occurs before it becomes hydrated.

For many G-proteins, a glutamine residue in switch II interacts with the water molecule and can be induced by the GAP to place the water molecule suitably for attack. Complexes of G-proteins with GDP and AlF_4^- have given detailed insights into the different states during GTP hydrolysis. Studies of the neighboring residues have led to the conclusion that the γ -phosphate abstracts a proton from the water molecule, leading to the attack. This type of mechanism is called substrate-induced catalysis and does not happen unless the water molecule is properly placed. The Arg-finger normally stabilizes the transition state. This can either be found in the GTPase itself (*cis*) or as part of the GAP (*trans*).

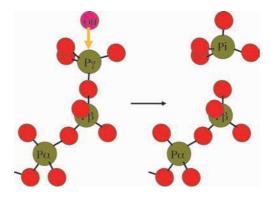


Fig. 8.28 The in-line sn² attack by a hydroxyl in on the γ -phosphate of an ATP or GTP molecule. The result can be compared to an umbrella turned inside out by a strong wind.

Fatty Acid Synthase — a Multifunctional Enzyme

Fatty acids are essential molecules in membranes and with many other roles in molecular systems. The synthesis of fatty acids is mostly done by fatty acid synthases (FAS). Several additional enzymes are involved in generating the substrates for this synthesis and some further enzymes proceed to use the main product, palmitate, for the production of specific molecules for energy storage or structural purposes. The synthesis occurs by a stepwise elongation of activated precursors two carbons at a time. In animals and fungi, the synthesis is carried out by an $\alpha_6\beta_6$ dodecamer with a molecular mass of 2.6 MDa. However, in mammals, the enzyme is a single polypeptide forming an α_2 dimer where each monomer has a molecular mass of 270 kDa. These are multidomain proteins that harbor all the different activities needed for the synthesis of fatty acids; they are called FAS type-I. These proteins are thus multifunctional enzymes. Bacteria and eukaryotic organelles, however, normally have the different activities on separate enzymes coded for by separate genes. This arrangement is called FAS type-II. The bacterial enzymes are functionally and structurally related to the different domains of the FAS type-I enzymes. The FAS type-1 belong to the group of multifunctional enzymes that perform iterative condensations of carboxylic acids (polyketide synthases) or amino acids (non-ribosomal peptide synthetases).

8.4.1 Structure

A central component in fatty acid synthesis is the acyl carrier protein (ACP), which carries reaction intermediates between the different active sites. This increases the local concentration of the substrate at the catalytic sites to allow efficient catalysis of fatty

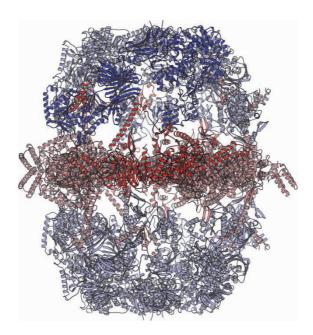


Fig. 8.29 • The structure of a fungal fatty acid synthase from *Thermomyces lanuginosus*. The central disc of the α subunits is shown in red and the doming β subunits are shown in blue. One α subunit and one β subunit are shown in darker colors (PDB: 2UV9 and 2UVA).

acids. The catalytic cycle is performed by substrate shuttling between different catalytic sites. Once the substrate reaches 16 to 18 carbon atoms in length it is released by a thioesterase (TE). High-resolution structures of individual bacterial subunits are known, as well as the complete enzymes from fungi and a medium-resolution structure from pigs.

The structures of FAS from the fungi *Saccharomyces cerevisiae* and *Thermomyces lanuginosus* (Figure 8.29) show a central wheel composed of six α subunits (also called FAS-2, 1878 amino acids), with upper and lower domes composed of three β subunits each, enclosing the reaction chambers. The β subunits (also called FAS-1, 2080 amino acids) are arranged as trimers with three-fold (C₃) symmetry while the six α subunits are arranged as dimers that also display three-fold symmetry (32 or D₃ symmetry (Figure 8.30). The α subunits are intertwined and their N-termini, three on each side of the central ring, form connections to the two domes each formed by the three β subunits. The particle is 270 Å in height by 250 Å in width and has a number of windows into the reaction chambers. There are also windows connecting the two reaction chambers.

The structure of the enzyme from pigs, composed of two identical 270 kDa chains, is very different. The general shape is an X with two legs and two arms (Figure 8.31). The dimensions are 210 x 180 x 90 $\rm \mathring{A}$ with an approximate two-fold axis running vertically.

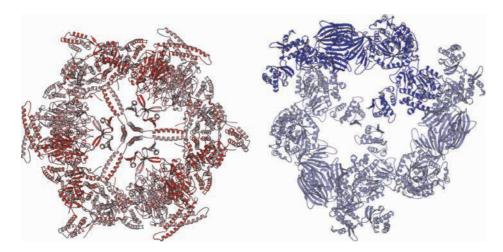


Fig. 8.30 ■ Left: The central disc of the fatty acid synthase is composed of six intertwined copies of the α subunits and has D₃ symmetry. Right: Three copies of the β subunit form each dome of the enzyme. Both are viewed down a three-fold axis. Three α subunits and one β subunit are highlighted.

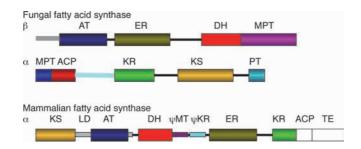
8.4.2 Enzymatic Function

8.4.2.1 Functional domains

The fungal enzyme has eight functional domains: ACP (acyl carrier protein), KR (ketoacyl reductase), KS (ketoacyl synthase), PT (phosphopantheine transferase) of the α subunit and AT (acetyltransferase), ER (enoyl reductase), DH (dehydratase), and MPT (malonyl/ palmitoyl transferase) of the β subunit. The MPT domain has components from the C-terminus of the β subunit and the N-terminus of the α subunit (Figure 8.31). In addition, the β subunit has four domains without catalytic activity. The mammalian single subunit has in the lower portion the domains KS and MAT (malonyl-acetyl transferase) involved in condensation reactions and in the upper portion the DH, ER, KR, ACP and TE (thioesterase) domains. In addition, there is a structured linker region (LD) on both sides of the MAT domain and a pseudomethyl transferase (\psi PT) and a pseudoketoacyl reductase (\psi KR) domain between the DH and ER domains.

8.4.2.2 Catalysis

The functional roles of the domains in the synthesis of fatty acids are shown in Figure 8.32. The AT and MPT domains of the β subunit have similar folds that are related to the fold of ferredoxin. Their active sites, containing histidine and serine residues, have been identified in deep crevices but are different with respect to their different substrates and functions. In particular, MPT has a conserved arginyl residue in the base of the catalytic



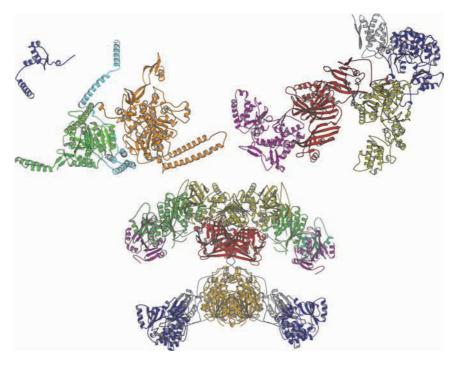


Fig. 8.31 Top: The organization of the functional domains in fatty acid synthases. *Middle:* The subunit structures of the fungal enzyme α (*left*, PDB: 2UV9) and β (*right*, PDB: 2UVA). The malonyl/palmitoyl transferase is composed of parts of both subunits of the fungal enzyme. In mammals, both the acetyl and malonyl transfer functions are performed by the MAT domain. *Below:* The structure of the mammalian dimeric enzyme with the shape of an X (PDB: 2VZ8). The lower part is the condensing part and the upper part is modifying. The ACP and TE domains are invisible due to flexibility.

pocket for the interaction with the carboxyl group of the malonyl moiety. The narrow catalytic cleft of AT prevents the binding of long-chain substrates, whereas the catalytic cleft of MPT is larger and hydrophobic to allow for binding of large substrates.

The ACP domain shuttles the substrate between the different active sites in the reaction chamber. In fungi, it is part of the α subunit and hinged on two flexible linkers (Figure 8.33). In a structure of *S. cerevisiae* fatty acid synthase it is stalled at the active site

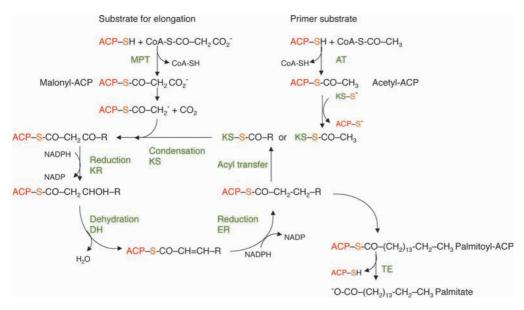


Fig. 8.32 ■ The main steps in fatty acid synthesis and the role of some of the catalytic domains (marked in green). The initial step in the synthesis of fatty acids is the transfer of an acetyl moiety from coenzyme A (CoA) to the ACP (red). The substrates are attached through a thioester bond to the terminal sulfur. The acetyl group is the two-carbon primer in fatty acid synthesis and is subsequently transferred to the active site cysteine of ketoacyl synthase (KS). Then ACP is charged with a malonyl moiety. This is decarboxylated, producing a highly reactive acetyl carbanion that reacts with the acetyl group or growing fatty acid attached to the KS cysteine. The substrate is subsequently shuttled to the ketoacyl reductase (KR) where NADPH is used for the reduction. The dehydratase (DH) eliminates a water molecule from the substrate, which is further reduced by the enoyl reductase (ER) and another molecule of NADPH. The substrate molecule is now two-carbon atoms longer and again becomes attached to the KS domain. The reaction proceeds until C_{16} or C_{18} fatty acids are transferred back to CoA by MPT. In mammals, there is also a transesterase domain (TE) that finally releases the substrate by hydrolysis.

of KS and getting visible. Yeast ACP is about 18 kD and completely helical. Its two linkers are attached to the center of the disc of the α subunits and the peripheral anchor attaches to the N-terminus of the α subunit. The 18 Å long phosphopantetheine (PPT) group is the chemical group to which the substrate is covalently bound. It is attached to a serine residue of ACP by the PT domain. The PPT group is located opposite two linkers of the ACP domain to allow it to reach a maximal distance.

KS catalyzes the condensation of the acyl and malonyl moieties. The KS domains are part of the α subunits and occur as dimers in the central wheel, but their active sites point into opposing reaction chambers. Its fold is related to thiolases. KS catalyzes all condensation steps. Thus, the acyl primer substrates bound to their active site cysteines can be short or long (C_2 to C_{16}).

The KR domain starts the reduction of the β -carbon using NADPH. It has a Rossmann fold. The DH domain eliminates water from hydroxyacyl-ACP. It has a long hydrophobic

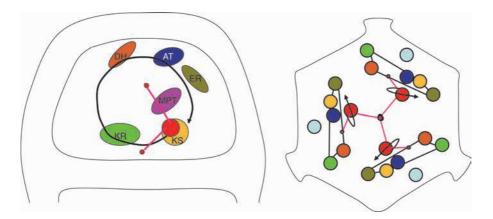


Fig. 8.33 \blacksquare The distribution of active sites within one reaction chamber of FAS. *Left:* The upper reaction chamber is seen as a cross-section. ACP with its two linkers are seen in red. The lower one originates from the center and the upper one (peripheral) originates from the N-terminus of the α subunit. The black arrow shows the orientation of the cyclic visits to four of the catalytic sites. *Right:* Birds-eye view of the reaction chamber showing all three ACP domains and the arrangements of the active sites that each of them visit.

tunnel to allow binding of long substrates. The ER domain is also dimeric. It catalyzes the final reduction step in the elongation cycle again using NADPH as reductant. It has a TIM barrel fold. The domain also has a flavin mononucleotide (FMN) bound at the base of the TIM barrel.

The active sites of six different catalytic domains (AT, MPT, KS, KR, DH and ER) are all oriented towards the inside of the two reaction chambers, while the PT domain is located on the outside in the fungal enzyme. Each reaction chamber contains three ACP domains. The arrangement of ACP and the active sites in the reaction chambers allow efficient catalysis of fatty acids. The flexible linkers of ACP make one set of the active sites easily accessible for the substrate. After substrate loading by AT or MPT, the ACP domain can visit the subsequent active sites in a clockwise manner (Figure 8.33) viewed from the inside of the reaction chamber. After such a cycle, the acyl chain is transferred (two-carbon atoms longer) back to the cysteine of KS so that ACP can pick up a new malonyl moiety from the MPT domain. Due to the double linkers, the ACP domain prefers to move essentially along a circle, which is in agreement with the placement of the different active sites (Figure 8.33). Each ACP domain interacts with active sites from two α and two β subunits.

The mammalian system with a very different design may function in a similar manner. The two ACP domains are again bound to flexible linkers and can access the different enzymatic sites. The linker between the top and bottom part could allow a rotation of 180° between them. In addition, the flexibility would allow the top to rock by upto 25° with regards to the bottom. Thus, the reaction chambers of the two ACP units would easily allow access to all different subsites.

For Further Reading (Sections 8.1 and 8.2)

Original Articles

Xu H, Faber C, Uchiki T, et al. (2006) Structures of eukaryotic ribonucleotide reductase I provide insight into dNTP regulation. Proc Natl Acad Sci USA 103: 4022–4027.

Xu Y, Feng L, Jeffrey PD, et al. (2008) Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms. Nature 452: 56-61.

Reviews

Eklund H, Uhlin U, Farnegardh M, et al. (2001) Structure and function of the radical enzyme ribonucleotide reductase. Prog Biophys Mol Biol 77: 177-268.

Jordan A, Reichard P. (1998) Ribonucleotide reductases. Ann Rev Biochem. 67: 71-98.

Krishnamurthy VM, et al. (2008) Carbonic anhydrase as a model for biophysical and physicalorganic studies of proteins and protein-ligand binding. Chem Rev 108, 946–1051.

Larsson KM, Jordan A, Eliasson R, et al. (2004) Structural mechanism of allosteric substrate specificity regulation in a ribonucleotide reductase. Nat Struct Mol Biol 11: 1142–1149.

Lindskog S, Liljas A. (1993) Carbonic anhydrase and the role of orientation in catalysis. Curr Opin Struct Biol 3: 915–920.

Nordlund P, Reichard P. (2006) Ribonucleotide reductases. Ann Rev Biochem 75: 681–706.

For Further Reading (Section 8.3)

Original Articles

Abrahams JP, Leslie AGW, Lutter R, Walker JE. (1994) Structure at 2.8Å resolution of F₁-ATPase from bovine heart mitochondria. Nature 370: 621–628.

Ariga T, Muneyuki E, Yoshida M. (2007) F1-ATPase rotates by an asymmetric, sequential mechanism using all three catalytic subunits. Nat Struct Mol Biol 14: 841–846.

Bourne HR, Sanders DA, McCormick F. (1991) The GTPase superfamily: Conserved structure and molecular mechanism. Nature 349: 117–127.

deVos A, Tong L, Milburn MV, et al. (1988) Three-dimensional structure of an oncogene protein: Catalytic domain of human c-H-ras p21. *Science* **232**: 1127–1132.

Kabaleeswaran V, Puri N, Walker JE, et al. (2006) Novel features of the rotatory catalytic mechanism revealed in the structure of yeast F1 ATPase. EMBO J. 25: 5433-5442.

Lupas AN, Martin J. (2002) AAA proteins. Curr Opin Struct Biol 12: 746–753.

Noji H, Yasuda R, Yoshida M, Kinosita K. Jr. (1997) Direct observation of the rotation of F1-ATPase. Nature 386: 299-302.

Pai EF, Kabsch W, Krengel U, et al. (1989) Structure of the guanine-nucleotide-binding domain of the Ha-ras oncogene product p21 in the triphosphate conformation. Nature 341: 209–214.

Pan X, Eathiraj S, Munson M, Lambright DG. (2016) TBC-domain GAPS for Rab GRPases accelerate GTP hydrolysis by dual-finger mechanism. Nature 442: 303–306.

Reviews

- Dimroth P, von Ballamos C, Meier T. (2006) Catalytical and mechanical cycles in F-ATP synthases. *EMBO Rep* **7**: 276–282.
- Gao YQ, Yang W, Karplus M. (2005) A structure-based model for the synthesis and hydrolysis of ATP by F1-ATPase. *Cell* **123**: 195–205.
- Iyer LM, Leipe DD, Koonin EV, Aravind L. (2004) Evolutionary history and higher order classification of AAA+ ATPases. *J Struct Biol* **146**: 11–31.
- Muench SP, Trinick J, Harrison MA. (2011) Structural divergence of the rotatory ATPases. *Quart Rev Biophys* **44**: 311–356.
- Stewart AG, Laming EM, Sobti M, Stock D. (2014) Rotatory ATPases Dynamic molecular machines. *Curr Opin Struct Biol* **25**: 40–48.
- Tucker PA, Sallai L. (2007) The AAA+ superfamily A myriad of motions. *Curr Opin Struct Biol* 17: 641–652.
- Weber J. (2007) ATP synthase the structure of the stator stalk. TIBS 32: 53–56.

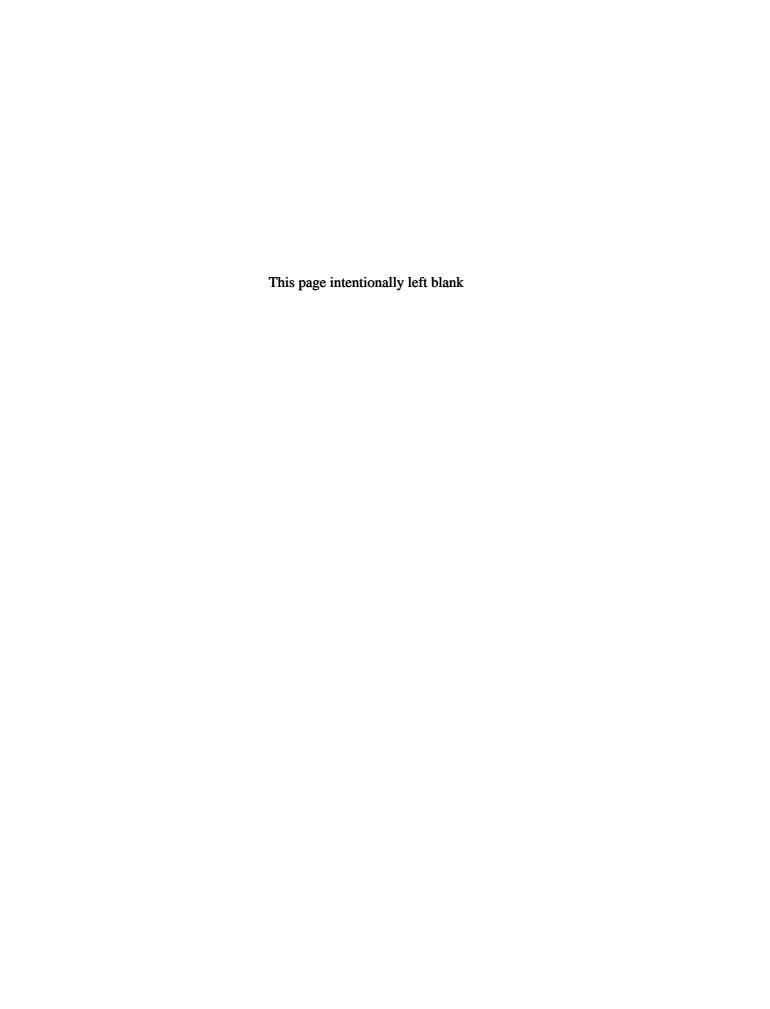
For Further Reading (Section 8.4)

Original Articles

- Jenni S, Leibundgut M, Boehringer D, et al. (2007) Structure of fungal fatty acid synthase and implications for iterative substrate shuttling. *Science* **316**: 254–261.
- Leibundgut M, Jenni S, Frick C, Ban N. (2007) Structural basis for substrate delivery by acyl carrier protein in yeast fatty acid synthase. *Science* **316**: 288–290.
- Lomakin IB, Xiong Y, Steitz TA. (2007) The crystal structure of fatty acid synthase, a cellular machine with eight active sites working together. *Cell* **129**: 319–332.
- Maier T, Leibundgut M, Ban N. (2008) The crystal structure of a mammalian fatty acid synthase. *Science* **321**: 1315–1322.

Reviews

- Leibundgut M, Maier T, Jenni S, Ban N. (2008) The multienzyme architecture of eukaryotic fatty acid synthase. *Curr Opin Struct Biol* **18**: 714-725.
- Xu W, Qiao K, Tang Y. (2013) Structural analysis of protein-protein interaction in type I polyketide polymerases. *Crit Rev Biochem Mol Biol* **48**: 98-122.



Genome Structure, DNA Replication and Recombination

9.1 Organization of the Genome

The double-stranded DNA in the genome of an organism needs to be organized to fit into the limited space available. This is done on several levels, particularly in eukaryotes. Here the complex between DNA and proteins is known as chromatin. The proteins are divided into: (i) histones and (ii) non-histone chromosomal proteins or non-histones, and they represent half of the total mass of the chromatin. If chromatin is subjected to treatments that cause its partial unfolding, it appears in the electron microscope as a series of beads on a string. The beads are called nucleosomes and the string connecting the beads is called linker DNA. The DNA wound up on histones form the nucleosomes. The nucleosomes are further organized into 30 nm fibers.

Each nucleosome consists of a segment of DNA, around 147 base pairs, and a protein octamer of two copies each of four histone proteins: H2A, H2B, H3 and H4 (Figure 9.1). The histones are very important both in the replication and transcription steps. Histone variants exist for H2A, H2B and H3. The double-stranded DNA is wound about 1.65 turns around the octamer in a left-handed superhelix (Figure 9.2, *left*). About half of the surface of the DNA is occluded, which makes it poorly accessible for replication, transcription and other DNA dependent activities. The amino acid sequences of histones are highly conserved between different species, and histones have long N-terminal tails. The histones are rich in arginine and lysine residues and their positive charges neutralize the negative charge from the DNA phosphate groups. Histone chaperones assist in forming the histone octamers and prevent unsuitable aggregation.

The four histones are proteins of 102–135 amino acid residues and share a common structural motif called the histone fold (Figure 9.2, *top right*). The core of the fold is composed of three α -helices connected by two loops. This structure is preserved in all eukaryotes.

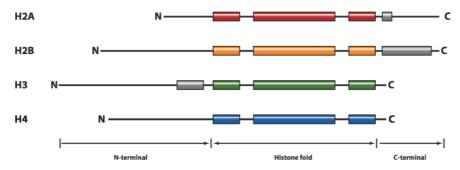


Fig. 9.1 ■ Schematic illustration of the organization of histone proteins. They all have a central region containing the histone fold and N- and C-terminal extensions.

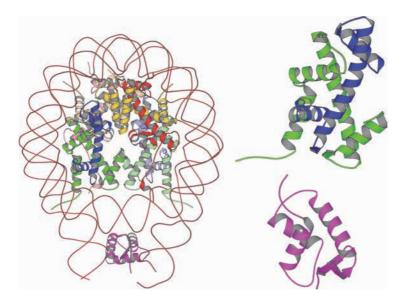


Fig. 9.2 ■ *Left*: The octameric complex of histone proteins forms the center and the DNA is wound around. The dyad axis goes vertically through the particle. The color scheme of the histone subunits in the core particle is the same as in Figure 9.1. A copy of histone H5 is bound at the bottom of the drawing. *Top right*: A dimer of histone proteins H3 (green) and H4 (blue). *Bottom right*: The "winged helix" structure of linker histone H5 (PDB: 4QLC).

There are hundreds of hydrogen bonds, salt linkages and hydrophobic interactions formed between the histone octamer and the corresponding DNA helix. The DNA structure is not uniformly wound, but several distortions and kinks are observed.

The long N-terminal tails extend beyond the nucleosome and can be methylated or acetylated in addition to several other modifications (see textbox, Chapter 10). These changes influence the packing of the nucleosomes and play important regulatory roles, for example, during gene expression. The pattern of modifications is recognized by other proteins, which can bind and regulate the transcription of certain genes. The combination of

histone modifications has been called the histone code and works as combinatorial signals that direct the binding of different molecules and subsequent events. When the DNA is replicated for a daughter cell, the structure of the chromatin is also copied. Thus, the histone code is also copied as part of what is called epigenetics. Epigenetics describes how dynamic modifications of DNA and histones can switch genes on or off. Thus, not only the genes in the DNA are important but also the extent to which these genes are accessible to transcription due to histone and DNA modifications.

The nucleosomes are packed into higher-level chromatin structures. Histone H1 or in some species H5, participates in this. Mammals have 11 subtypes of H1 histones with globular domains of about 80 amino acid residues and short N-terminal (20–35 residues) and long C-terminal (100 residues) tails. In the C-terminal tail, upto half of the amino acid residues are basic, primarily lysines. The fold of the globular domain is different from the other histones (Figure 9.2). Histones H1 or H5, which are present in about one copy per nucleosome, are called linker histones. They bind at the two-fold axis and interact with the DNA that enters and exits the nucleosome. The tails bind to linker DNA and participate in the formation of higher-order structures. Nuclease cleavage of nucleosomes with bound linker histones shows a protection of around 168 base pairs of DNA. The extra base pairs are symmetrically distributed on either side of the nucleosome.

The first level of compacting the nucleosomes is the 30-nm structure. The structure of tetranucleosomes has been investigated by crystallography and larger segments have been studied by cryo-EM. Groups of four nucleosomes are arranged in a unique way (Figure 9.3). These tetranucleosomes are the building blocks in forming a 30-nm fiber. The structure is composed of two stacks of nucleosomes rather than a single one (Figure 9.4). The connection between successive nucleosomes goes across the interior of the fiber. The H1 histone is located asymmetrically at the dyad axis. The C-terminal tail of histone H1 plays important roles in the organization of the DNA linker.

The 30-nm structure still needs to be condensed more than 100 times to reach the in vivo chromosome structure. This higher-order packaging and the global structure of chromosomes remain poorly understood.

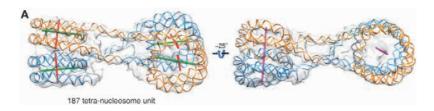


Fig. 9.3 ■ The organization of the DNA in the tetranucleosome in the 30-nm structure. The two brown nucleosomes are oriented 90° apart and stacked on top on another pair with the same arrangement. (Reproduced with permission from Song F, et al. (2014) Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. Science 344: 376-380. Copyright (2014) AAAS.)

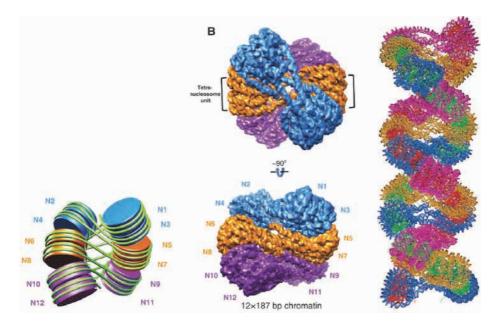


Fig. 9.4 ■ Left: Twelve nucleosomes packaged as three tetranucleosomes seen schematically from the side. Middle: Twelve nucleosomes seen from top and from the side. Right: A fragment of the 30-nm structure. (Reproduced with permission from Song F, et al. (2014) Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. Science 344: 376-380. Copyright (2014) AAAS.)

Replication

Before a cell can divide, its genome must be duplicated. This is done by DNA polymerases assisted by a large number of other proteins (Table 9.1). Metabolism of DNA is one of the central activities in each cell and includes, in addition to replication also repair, recombination and degradation. These processes help to transmit the genetic information from one cell to the next and from one generation to another, but we will focus only on a few of them.

Matthew Meselson and Franklin Stahl demonstrated that both parental strands in the initial DNA double helix serve as templates for new strands. The two resulting DNA double helices each have one old and one new strand. This is called semi-conservative replication.

The replication in eukaryotes takes place in at least four steps: (i) a segment of DNA from the highly organized nucleosomes is liberated from histones and other proteins that are compacting its structure, (ii) the DNA double helix is unwound and the two strands are separated and therefore made available for new base pairing, (iii) new nucleotides

| Function | E. coli | Archaea | Eukaryotes |
|----------------------------|--|----------------|---------------------------------|
| Origin recognition | DnaA | Cdc6/Orc1 | ORC, Cdc6, Cdt1 |
| Helicase | DnaB | Cdc6/Orc1, MCM | MCM (Mcm2, 3, 4, 5, 6) |
| Helicase loader | DnaC, PriA | GINS | GINS, Cdc45 |
| ssDNA-binding protein | SSB | | RPA (p14, p32, p70) |
| Primer synthesis | DnaG | DNA primase | Pol α |
| Clamp loader | γ complex $(\tau_3 \delta \delta' \chi \psi)$ | RFC | RF-C (p140, p40, p38, p37, p36) |
| Sliding clamp | β | PCNA | PCNA |
| Replicative DNA polymerase | Pol III $(\alpha, \varepsilon, \theta)$ | PolB, PolD | ΡοΙ α, δ, ε |
| 5' to 3' exonuclease | Pol I N-domain | Fen1, Dna2 | Fen1, Dna2 |
| DNA ligase 1 | DNA ligase | DNA ligase | DNA ligase 1 |

TABLE 9.1 Names of Proteins Involved In DNA Replication

are linked by covalent bonding and the growing strand sequence made through complementary base pairing with the template strand and (iv) histones are added back to both duplexes with appropriate modifications to retain the epigenetic signals.

9.2.1 Histone Chaperones

The removal and adding back of histones during replication is a complex and essential process. During replication, the histone modifications, associated with some specific part of the genome, should be retained for both DNA duplexes. Histone chaperones participate in this process to escort the histones through replication, DNA repair and transcription.

There is a wide range of histone chaperones. Several chaperones interact with different variants of H3 in complex with H4. Others interact with H2A-H2B. For example, ASF1 (Figure 9.5), which binds disrupted histones ahead of the replication fork, will after the fork provide histones for the assembly of new nucleosomes. Other parts of the ASF1 structure participate in interactions with histone deposition factor proteins, which also are active during DNA synthesis.

Ahead of the replication fork the histones are removed and ASF1 acts as a histone acceptor, binding primarily to the H3-H4 dimer. Recycling and deposition of newly synthesized histones is coordinated and ASF1 may participate in this. The complex interplay remains to be established, and also the details of how the H2A-H2B dimers are handled. The field of histone chaperones is of significant interest since mutations or expression levels can lead to severe diseases including cancer.



Fig. 9.5 ■ The histone chaperone ASF1 (brown) binds a heterodimer of H3–H4 on the side where H3 binds H3' thereby preventing tetramerization (PDB: 2HUE).

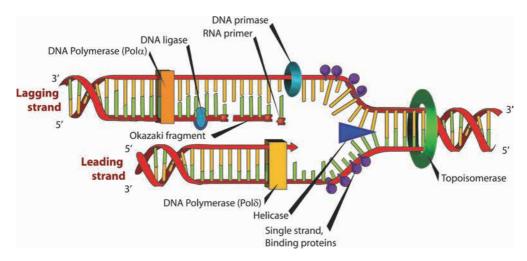


Fig. 9.6 ■ An overview of some of the components involved in replication of DNA in eukaryotes. Drawing by Mariana Ruiz Villareal deposited at Wikipedia.

9.2.2 DNA Synthesis

In DNA replication, the DNA polymerases work in the direction from the 3' to 5' end of the strand copied and the newly synthesized chain begins at 5' and ends at 3'. The double-stranded DNA has to be split into two template strands. This leads to a branch point in the DNA called the replication fork (Figure 9.6). One strand goes in the direction 3' to 5'. This strand is the leading strand where the replication fork moves further and further. The other strand, the lagging

strand, going in the 5' to 3' direction, is also replicated in the 3' to 5' direction. This is done in the form of fragments starting from the moving replication fork in one position upto the 3' end of the previous fragment. Thus, the lagging strand will be replicated as fragments, called Okazaki fragments, which will be linked together by a DNA ligase.

9.2.2.1 Recognition of the origin of replication

Before a cell can divide, the origin for replication has to be located. In the cell, each separate DNA molecule (e.g. a chromosome) has at least one starting point for replication. In bacterial chromosomes, there is usually only one origin of replication, but linear eukaryotic chromosomes may have many origins of replication. The origins of replication are relatively rich in A-T base pairs, which are more easily opened than G-Cs. Several origins can be used simultaneously and result in many replication forks.

In bacteria, 10–20 copies of the monomeric protein DnaA oligomerize at the origin of replication. This causes part of the DNA to melt and DnaC helps to load the hexameric helicase DnaB onto the melted region of DNA.

In eukaryotes, the origin recognition complex (ORC) is composed of six different subunits (Orc1-Orc6) and protein Cdc6 (Figure 9.7). Subunit Orc1 and the Cdc6 protein have considerable sequence similarity. These proteins, except Orc6, belong to the AAA+ class of ATPases (see Section 8.3). In archaea, Orc1 and Cdc6 alone manage the origin recognition and Orc1 is shaped like a C that binds to the origin DNA like the claws of a lobster and bends the DNA. In eukaryotes, the ORC complex also has a crescent shape.

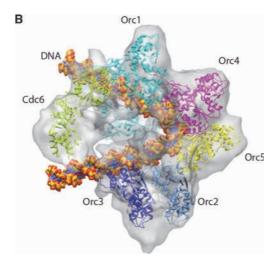


Fig. 9.7 ■ A model of the origin recognition complex (ORC) and Cdc6 from yeast bound to 72 base pairs of DNA. The Orc6 subunit is probably placed at the sharp turn of the DNA. (Reproduced with permission from Sun et al. (2012) Cdc6-induced conformational changes in ORC bound to origin DNA revealed by cryo-electron microscopy. Structure 20: 534–544. Copyright (2012) Elsevier.)

How the ATP hydrolysis is orchestrated in relation to the conformational rearrangements remains unknown.

9.2.2.2 The replisome

A huge protein aggregate, the replication complex or the replisome, is central in DNA replication and composed of a helicase with its loader, a DNA primase, a sliding clamp with its loader and a DNA polymerase. The helicase binds at the origin of replication and induces a local denaturation site by breaking the double helical structure. A DNA polymerase cannot start to synthesize a polymer strand on its own, but it needs a helping DNA or rRNA, also called a primer. Usually the primer is a short strand of rRNA, which has been presynthesized by an rRNA polymerase called primase. DNA is then replicated in both directions by two moving replication forks. Complementary-base pairing guides the synthesis of the new strands.

9.2.2.3 DNA helicases and helicase loaders

Upon recognition of the origin of replication, the DNA helicase breaks the forces that hold the two DNA strands together. Helicases are molecular motors that couple the energy of ATP hydrolysis to the unwinding of DNA (or rRNA) double helices. The helicases are essential not only for replication, but also for recombination, repair, transcription and translation. Helicases can be grouped into different families based on the conservation level of characteristic motifs and by the direction of translocation along the nucleic acid template (Table 9.2). Their motor domains all belong to the P-loop-containing ATPases,

Super-Fold of ATPase **Family** Oligomer and Helicase Part **Examples of Other Members** E. coli Enzymes SF₁ 2 x RecA domain Monomer UvrD, Rep (DNA PcrA (bacterial repair) recombination) SF₂ DEAD-box rRNA helicases, one Monomer 2 x RecA domain RecQ, UvrB (DNA subunit of transcription factor repair) TFIIH SF3 Hexamer AAA + domain SF₄ Hexamer RecA domain DnaB (DNA replication) SF 5 Hexamer RecA domain Rho (transcription termination) SF₆ AAA + domainHexamer Mcm2-7 (DNA replication in eukaryotes)

TABLE 9.2 Families of DNA Helicases

but they mostly have several other domains of various functions. Single strand-binding proteins assist the DNA helicase to stabilize the open structure. Also other enzymes, such as DNA topoisomerases, assist the moving replication forks (Section 9.2.3).

SF1 and SF2 helicases are monomers with two RecA-like domains (Section 8.3.1), both of which have inserted domains. The RecA-like domains (Figure 9.8) are the cores of the helicase motor domain. Several conserved sequence motifs are involved in coupling the energy of nucleotide hydrolysis to the physical separation of the two complementary DNA and/or rRNA strands. Some of these motifs, including the P-loop motif (GXXXXGKS) create the binding site for ATP. The ATP is bound between the two RecA domains and the ssDNA moves in a tunnel, which has insufficient room for a dsDNA. In essence, conformational changes between the protein domains due to ATP binding, hydrolysis and dissociation moves the double-stranded DNA into contact with a negatively charged area of the protein that forces the DNA strands to separate.

The SF2 helicases are found in all organisms and are mostly used in DNA repair. RecQ, for example, is a monomer with three conserved domains: (i) helicase with two RecA-like subdomains, (ii) RecQ C-terminal (RQC) and (iii) helicase-and-RNaseD-like-C-terminal

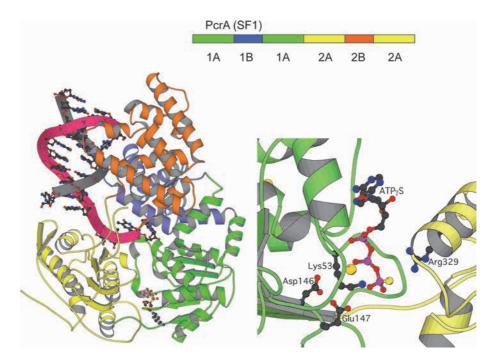


Fig. 9.8 ■ Top: The domain arrangement of PcrA (SF1) with the two RecA-like domains (1A and 2A) which each has an insert (2A and 2B). Bottom left: PcrA with a tailed duplex of DNA. The ATP is bound between the RecA-like domains 1A and 2A and the ssDNA goes through a narrow tunnel in the protein. Bottom right: The ATP binding site and the residues involved in catalysis. Lys53 is from the P-loop and Asp146 and Glu147 are from the second conserved motif (DEXH) of the first RecA-like domain. Arg329 from the other domain is close to the active site, similar to arginine fingers in other P-loop-containing proteins (PDB: 10YY).

(HRDC). The RQC and HRDC domains are involved in substrate specificity and targeting as well as in interactions with other proteins of the replication complex. A crystal structure of the RQC domain, which has a winged-helix motif, in complex with dsDNA shows the binding to 8 terminal base pairs. A prominent β -hairpin, part of the β -wing of the winged-helix motif, has a function like a knife that separates the two strands of DNA. Aromatic side chains of the hairpin stack both with paired and unpaired bases of the DNA. The ATP binding, hydrolysis and release is again converted into conformational changes between the RecA-like domains that leads to gripping and releasing of the ssDNA, leading to its translocation.

The hexameric helicases (SF3-6; Table 9.2 and Figure 9.9) are ring structures and have either RecA (SF4-5) or AAA+ like (SF3 and SF6) ATPase domains. While SF3 and SF6 have

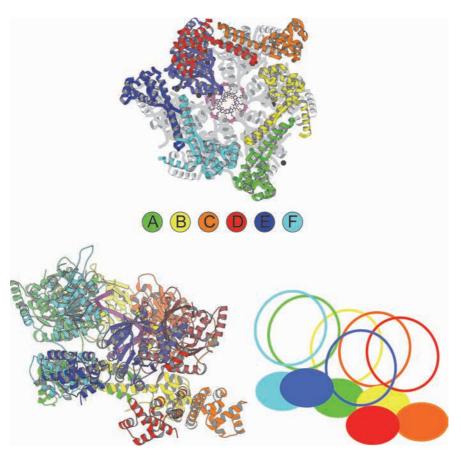


Fig. 9.9 ■ The hexameric structure of the bacterial replicative helicase DnaB (SF4) with ssDNA bound in the inner cavity displayed in two orthogonal views. The subunits are displayed with different color (PDB 4ESV). *Top*: The six subunits are shown along the symmetry axis with the DNA in the middle. Only for the blue subunit both the N- and C-terminal domains are shown in color. *Middle*: The sequence of circles shows the order of the subunits. *Bottom left*: The side view in atomic detail. The three subunits closes to the viewer are shown fully in color. The N-termini are at the bottom and the C-termini at the top. *Right*: A schematic drawing showing the positions of the N-termini fully colored and the C-termini as an outline. The helical staircase structure has the gap between the blue and the red subunits.

flat ring structures with one DNA nucleotide bound per subunit, SF4 and SF5 form righthanded staircase structures with two nucleotides per subunit.

In bacteria, the protein DnaA first of all binds to the origin of replication. The helicase DnaB (SF4) is built of two domains where the larger CTD has a single RecA fold, binds ATP and has the role of translocating the helicase. The unliganded DnaB₆ has a flat double ring structure. A helicase loader loads the helicase onto the DNA by opening the closed ring of the helicases. The helicase loader in E. coli, DnaC, in complex with ATP binds to the DnaB to form the complex DnaB₆-DnaC₆.

The helicase loader protein is also built of two domains. Here the CTD has the AAA+ fold and binds ATP and also to the ssDNA. The NTD has a zinc-binding fold and interacts with the CTD of the helicase. The protein alone is a mixture of monomers and dimers. It binds to DnaB as three dimers. One dimer has a somewhat different orientation, which leads to a cleft in the structure, an open ring structure. The complex of the helicase with the helicase loader can form a spiral staircase structure that is probably needed for the loading of ssDNA (Figure 9.10). The complex is loaded onto the single-stranded DNA formed at the origin of replication by contact with DnaA. The helicase loader dissociates coincident with the formation of the staircase structure.

The structure of DnaB binding to ssDNA with ATP analogues is arranged like two lock-washers stacked on top of each other (Figure 9.9). The top layer of the NTD domains is organized like a trimer of dimers, which surrounds the ssDNA and can also bind the primase. In the bottom layer, the CTDs are organized into a pseudo-six-fold helical symmetry with a rotation of 60° between the subunits. The CTDs bind to the phosphates of the ssDNA. The NTD and the CTD of the same subunit generally make little contact with each other but are important to keep the covalent contacts. The linker region between the top and bottom subunit keeps the structure as a closed assembly. Eleven nucleotides interact with the six subunits but only 5 subunits contain the ATP analog, since the ATP-binding site between the top and bottom subunits is empty.

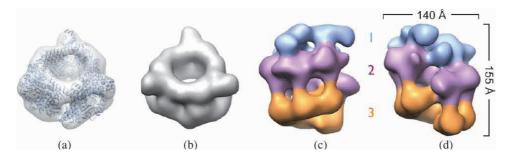


Fig. 9.10 ■ (a) and (b) The structure of the closed structure of the helicase DnaB. The upper level is due to the DnaB NTD arranged as a trimer of dimers and the lower level is the CTD arranged with a pseudo-six-fold symmetry. (c) and (d) The binding of DnaC (orange) to the DnaB CTD gives the structure a right-handed spiral conformation with a gap [panel (d)] through which the ssDNA can be inserted. (Reproduced with permission from Arias-Palomo E, et al. (2013) The bacterial DnaC helicase loader is a DnaB ring breaker. Cell 153: 438–448. Copyright (2013) Elsevier.)

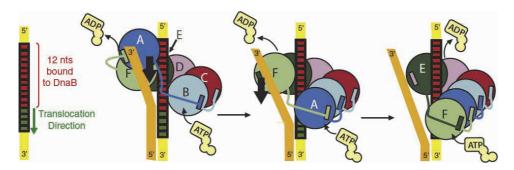


Fig. 9.11 ■ The movement of the helicase DnaB along the double-stranded DNA. In the first step (first arrow) subunit A, that has lost its ADP molecule, now picks up an ATP molecule and connects to the B subunit. In the second step subunit F undergoes the same procedure (Reproduced with permission from Itsathitphalsam O, et al. (2012) The hexameric helicase DnaB adopts a nonplanar conformation during translocation. Cell 151: 267–277. Copyright Elsevier.)

The mechanism of translocation of the helicase along the dsDNA may be a hand-overhand process in a manner like a rope climber (Figure 9.11).

Replication in archaea and eukaryotes is initiated at replicative origins by a complex of the six ORC subunits and Cdc6, which also function as the helicase loader. This complex first recruits a heptamer of Mcm2-7 (MCM) with Cdt1 to the origin of replication. After ATP hydrolysis by Orc1 and Cdc6, Cdt1 is released and subsequently a second heptameric complex is bound. In the double hexamer, the MCM rings interact through their zincbinding N-termini, while the C-termini contain the AAA+ ATPase motifs. Subsequently, the prereplicative complex disassembles and additional proteins are recruited. These are the heterotetrameric GINS complex and Cdc45 forming the 11-membered CMG complex. All components are essential for function and are conserved. MCM is the motor. The helicase now becomes active and replication progresses bidirectionally from the origin with one helicase hexamer as part of the replisome progression complex.

Further studies by cryo-EM of CMG bound to dsDNA with a single-stranded 3' overhang at low resolution, combined with components at higher resolution, have provided further insights (Figure 9.12). Here, the C-termini of the Mcm2-7 subunits no longer form a circle but are broken between subunits Mcm2 and Mcm5 to form a right-handed spiral. The GINS and Cdc45 subunits bridge the junction between the ends of the spiral. However, the N-termini of MCM retain the planar arrangement. This complex binds to the single-stranded part of the DNA and the C-terminal AAA+ parts of MCM are closest to the DNA, suggesting that the ATPase motor of the CMG complex leads the movement along and opening of the dsDNA. Cdc45 interacts with the leading strand of the forked DNA substrate while the lagging strand passes through the central hole between the MCM subunits.

SF4/SF5 with their RecA-like domains are related to the F₁-ATPase (Section 8.3.1). The subunits go through three states during the catalytic cycle, the ATP, ADP and empty

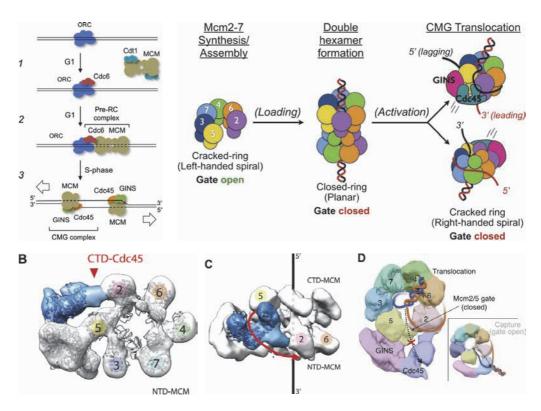


Fig. 9.12 ■ Top left: The initial steps of replication. The origin recognition complex (ORC) binds to the origin of replication. Cdc6 and Cdt1 bind to ORC before a double hexamer of MCM can associate with the ORC. For the initiation of replication ORC is disassembled and Cdc45 and GINS bind to MCM. The hexameric rings of the helicase can then progress in both directions of the DNA. (Reproduced with permission from Onesti S, MacNeill SA. (2013) Structure and evolutionary origins of the CMG complex. Chromosoma 122: 47-53. Copyright Springer Verlag GmbH.) Top right: The inactive Mcm2-7 is a cracked left-handed spiral. The crack occurs between Mcm2 and Mcm5. By the action of ORC and assisting proteins MCM forms a double hexamer closed around the double-stranded DNA at the origin. In the activated state, by the binding of GINS and Cdc45, the crack between Mcm2 and Mcm5 reopens but now forms a right-handed spiral where the activating proteins form a bridge. Here, the leading strand of the DNA interacts on the outside with Cdc45 while the lagging strand passes through the central hole in the ring of MCM subunits. *Bottom*: The cryo-EM densities of CMG seen in two orientations and an illustration of the path of the DNA. (Both figures were reproduced from Costa A, et al. (2014) DNA-binding polarity, dimerization, and ATPase ring remodeling in the CMG helicase of the eukaryotic replisome. eLife 3: e03273. Copyright Costa A, et al.).

states. Different from the F_1 -ATPase, these helicases have only one type of subunit. Thus, all six subunits of the hexamer participate in ATP hydrolysis. All subunits can contact the DNA at some stage.

Generally, all the helicases SF3-6 adopt three different conformations, possibly with three subunits in the ATP-binding conformation, two in the ADP-binding conformation

9.2.3 Topoisomerases

9.2.3.1 Roles and types

During replication, transcription and other cellular activities the strands of DNA are separated or untangled. The DNA and rRNA polymerases (that translocate within a DNA bubble) force the DNA to become supercoiled around its helical axis, both in a positive or a negative sense. Behind the polymerases, the DNA will be supercoiled in a negative sense. Negative supercoiling favors strand separation while positive supercoiling prevents it. The torsional stress caused by positive supercoiling can inhibit further polymerase activity. DNA topoisomerases are then needed to release this topological stress. These enzymes cut one (type I) or both strands (type II) of DNA to release the tense DNA topology and subsequently religate the DNA. The catalytic process involves binding one piece of DNA, called the G-segment, that is subsequently cleaved. The other segment of DNA, called the T-segment, can be transported through the G-segment.

Tyrosines are central in the catalytic process and form covalent links with the cleaved DNA. In the case of topoisomerase I, the tyrosine can either link with the 5'-end (type IA) or the 3'-end (type IB). After DNA cleavage, in type IA the non-cleaved strand of DNA is transported through the gap in the other strand. Type IB topoisomerases allow for controlled rotation of the free 5'-end of the DNA. In type II topoisomerases, tyrosine residues connect with both 5'-ends of the cleaved DNA. Here a dsDNA passes through the gap of the cleaved segment in an ATP-dependent reaction. Type II is also grouped into two subclasses, A and B, based on structural differences.

Topoisomerase poisons that inhibit these enzymes have been used extensively in antibacterial and anticancer therapy and remain of significant interest for developing new drugs.

9.2.3.2 Topo IA, single-strand DNA passage enzymes

Topoisomerase type IA is a monomer where the N-terminal region is composed of four domains arranged in a toroidal fashion, which can close around a dsDNA (Figure 9.13). The C-terminal region is not conserved.

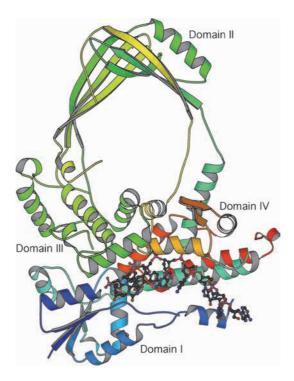


Fig. 9.13 ■ Topoisomerase IA. The four domains of the N-terminal region colored from N- to C-terminus: Blue to red (PDB: 3PX7).

Topo IA changes the topology of negatively supercoiled or underwound DNA without consumption of ATP. Domain I contains the active site and the enzyme binds to a single-stranded piece of DNA. The active site undergoes significant conformational changes in this process. The ssDNA is cleaved and its 5'-end becomes covalently attached to the active site tyrosine (domain III). The catalytic activity of Topo IA, like that of Topo II depends on two metal ions. The toroidal structure can open up to temporarily sequester the DNA strand that passes through the opened strand.

9.2.3.3 Topo IB, swivelases

Topoisomerases type IB can relax both negatively and positively supercoiled DNA without using ATP. Topo IB is also monomeric but entirely differently constructed (Figure 9.14). The fold is common but the size of the molecule is highly variable: 36kDa in bacteria and viruses, and 90kDa in eukaryotes. The enzyme has the shape of a C, which can clamp the DNA.

The active site tyrosine in Topo IB makes a transient covalent interaction with the 3'-end of the cleaved strand of DNA. The torsional strain of the DNA will drive the swiveling of the DNA until it is relaxed.

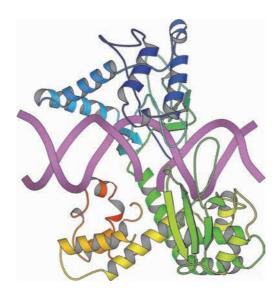


Fig. 9.14 ■ The structure of human topoisomerase IB with a bound dsDNA (PDB: 1A31).

9.2.3.4 Topo II, duplex DNA passage enzymes

Topoisomerases IIA and IIB catalyze the passage of one DNA duplex across another and have related constructions. The process is rather complex and is catalyzed by a dimeric enzyme. In Topo IIA, the N-terminal domains contain the site for binding and hydrolysis of DNA. One ATP is hydrolyzed to form the gate in the DNA for the passage of the other dsDNA and a second ATP is hydrolyzed to get the enzyme back into its initial state.

The enzyme in bacteria has two different subunits forming a heterotetramer, but in eukaryotes all domains are within the same dimeric polypeptide (Figure 9.15). The N-terminal part of the enzyme has three different domains, the ATPase domain of the GHKL family, the transducer domain and the topoisomerase-primase (TOPRIM) domain. The C-terminal side has a CAP domain containing the catalytic tyrosine and a C-terminal domain. The enzyme has three separate gates that are central for the process.

The G-segment binds with an U-shape (150° bend) to the binding groove across the middle part of the TOPO IIA dimer, built up of the TOPRIM domain, the winged-helix domain (WHD) and the tower domain. The bend in the DNA is caused by a conserved intercalating isoleucine in the wing-2 region of the WHD. Subsequently, a helix-turn-helix motif in WHD containing the catalytic tyrosine binds. The DNA binding to a topoisomerase does not need any specificity and does not primarily bind to the base pairs.

The T-segment of DNA can now access through the N-gate to its binding site. The enzyme binds ATP to both GHKL domains, which leads to a closure of the gate and cleavage of the G-segment. The spatial arrangement of the two tyrosines leads to a staggered cleavage of the G-segment four bases apart. Associated with hydrolysis of the ATP molecules,

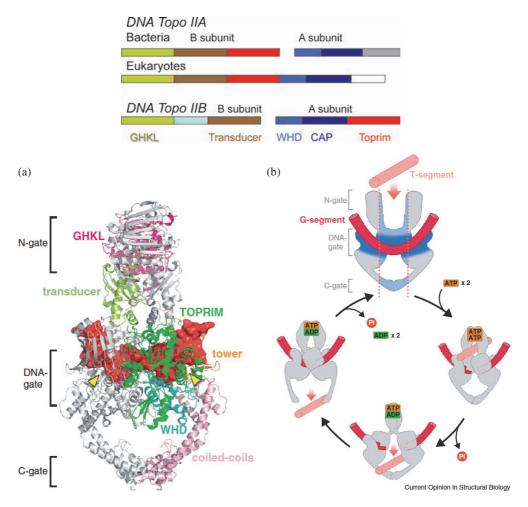


Fig. 9.15 ■ *Top*: The domain arrangements in bacterial and eukaryotic topoisomerases IIA and IIB. Bottom: The structure (left, PDB: 4GFH) and an outline of the catalytic process for topo IIA (right). Reproduced with permission from Chang C-C, et al. (2013). New insights into DNA-binding by type IIA topoisomerases. Curr Opin Struct Biol 23: 125-133. Copyright (2013) Elsevier.)

the T-segment is first able to pass through the cleaved G-segment and the middle gate, and subsequently through the third gate while the G-segment is reunited.

9.2.3.5 Sliding clamp and clamp loader

In replication, the polymerase needs to repeat the same catalytic function thousands of times. The binding of the DNA polymerase to the DNA is not strong and the enzyme can easily fall off. Therefore, the enzyme is kept associated with the template by a universal

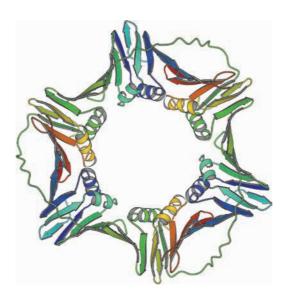


Fig. 9.16 ■ A sliding clamp: human PCNA (proliferating cell nuclear antigen) that encloses the newly replicated DNA (PDB: 1U7B).

mechanism. A processivity factor, a sliding clamp protein, is used to increase the rate of DNA synthesis by upto two orders of magnitude. These clamps form a ring that encloses the nucleic acid (Figure 9.16). The DNA polymerase is more strongly attached to the clamp than to the DNA. The sliding clamp is built of six repeating domains. In bacteria, it is built of a homodimer, but a homotrimer in eukaryotes. The domain has positively charged α -helices facing the DNA and β -structures on the outside.

In order to load the clamp onto the dsDNA, a specialized protein called the clamp loader is engaged. Its function is similar to the helicase loader (Section 9.2.2.3). The clamp loader is a pentameric enzyme (Figure 9.17). Sliding clamps need to be attached to the DNA both at the origin of replication on the leading strand and on each of the Okazaki fragments of the lagging strand. The clamp loader belongs to the AAA+ family of enzymes. With ATP, the loader binds to the clamp and breaks the ring open. The structure of the clamp loader forms a right-handed spiral that matches the helical symmetry of the DNA. It forces the sliding clamp into a right-handed open lock-washer structure. When the complex binds to primer-template DNA, the ATPase activity is induced. This changes the structure of the clamp loader to lose its affinity for the clamp. The clamp will then close around the DNA, bind the DNA polymerase and replication will start.

At this stage, the replisome is being assembled. The DNA, the helicase around the lagging strand, the primase making short rRNA primers, the DNA polymerase, the clamp and the clamp loader are all in place. In bacteria, the CTD of the three γ subunits (sometimes also called τ , Figure 9.17 bottom) can bind both to the DNA polymerase and the helicase to hold the replisome together.

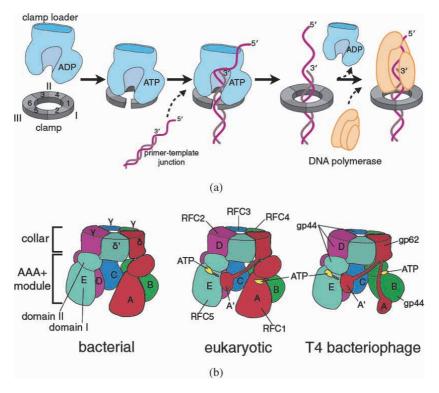


Fig. 9.17 ■ Top: The mechanism of clamp opening and insertion of the dsDNA by the clamp loader. Bottom: The arrangement of the clamp loader in bacteria, eukaryotes and T4 bacteriophage. (Reproduced from Kelch BA, et al. (2012) Clamp loader ATPases and the evolution of the DNA replication machinery. BMC Biol 10: 34–48. Copyright Kelch BA, et al.).

Bacteriophages, like T4, encode all proteins needed for its DNA replication including helicase, helicase loading protein, primase, sliding clamp and clamp loader proteins as well as DNA polymerase, an exonuclease to remove rRNA primers and a ligase to ligate lagging strand nicks.

9.2.4 DNA Polymerases

9.2.4.1 *Families*

The fitness of an organism depends on accurate replication and how it maintains its hereditary material, but also how to "permit" a low frequency of errors. In general, an error is generated only every 10⁹–10¹⁰ bases replicated. The speed of replication is upto 1000 nucleotides per second. This remarkable accuracy is achieved by precise nucleotide incorporation and an ability to remove wrongly inserted nucleotides. DNA polymerases

are the key enzymes in assuring the accuracy in replication, repair and recombination of DNA. Apart from the polymerization ability, many DNA polymerases also exhibit exonuclease (5'-3') and 3'-5' activities.

Since the discovery of the first polymerase activity in 1956 by A. Kornberg and coworkers, the number of known DNA polymerases has grown. It seems that nature created safety mechanisms by employing different polymerases for similar tasks. For example, normal DNA replication in eukaryotes requires three polymerases, Pol α , δ and ϵ , while other polymerases are involved in DNA repair or other activities needed for the handling of damaged DNA. Based on the primary sequence homologies and crystal structure analysis, DNA polymerases can be grouped into seven different classes: A, B, C, D, X, Y and RT (Table 9.3). The eukaryotic replicative polymerases α , δ and ϵ belong to class B. While six families use DNA as the template, reverse transcriptase (RT) polymerases can convert a single-stranded viral rRNA genome into double-stranded DNA incorporated into the host genome (provirus). Telomerases (Section 9.3) also belong to the RT family. In general, classes A, B, X, Y and RT follow the same general polymerization theme.

TABLE 9.3 Examples of DNA Polymerases Found In Different Species

| Class ^a | E. coli | $\textbf{Eukaryotes}^{\text{b}}$ | T4 Bacteriophage | Main Function |
|--------------------|----------------------------|----------------------------------|-------------------------|---|
| A | Pol I (Klenow fragment) | | | DNA repair |
| | | Pol γ | | Replication of mitochondrial DNA |
| В | Pol II | | | DNA repair |
| | DnaG | Pol α (4 su) | gp61 | Primase |
| | | Pol δ (4 su) | | Main replicase (lagging strand) |
| | | Pol ϵ (4 su) | | Main replicase (leading strand) |
| | | | gp43 | Replication of leading and lag- ging strand |
| С | Pol III (15 su) | | | Replication of leading and lag- ging strand |
| D | Pol D (2 su) | | | Replication in Archaea |
| X | | Pol β | | DNA repair |
| Y | DinB | | | Bypass synthesis |
| | | Pol η | | Bypass synthesis |
| RT | | | | Reverse transcriptase. rRNA dependent DNA synthesis |

^aThe classes are based on sequence similarity. Class C, to which the main bacterial replicase belongs, has no member in eukaryotes. The number of subunits are given in brackets.

^bSeveral more polymerases are found in mammals.

9.2.4.2 Structures of DNA polymerases

Once the proteins involved in compacting the DNA have been removed from a stretch of DNA, the helicases and topoisomerases assisted by a number of other proteins will separate the two strands of the duplex DNA and set the stage for the DNA polymerases. Both strands are replicated from a 3' towards a 5' location on the parent strand.

The basic architecture of DNA polymerases can be best illustrated by the first characterized polymerase, E. coli polymerase I (Pol I). This enzyme has a size of 109 kDa and has three domains. The C-terminal domain is the polymerase and the 5'-3' and 3'-5' exonuclease activities are performed by the two N-terminal domains. The C-terminal portion, lacking the 5'-3' exonuclease (Exo) domain, is called the Klenow fragment and was the first DNA polymerase structure to be solved. The basic structural organization is described as a right hand with fingers, palm and thumb, and is found in almost every DNA polymerase (Figure 9.18). The active site is located in the palm sub-domain, which forms the base of the crevice surrounded by the fingers and thumb sub-domains. The thumb part binds DNA and the fingers are important in the recognition and binding of nucleotides.

Structures have been determined for most types of DNA polymerases (Figure 9.19). Polymerases in classes A, B and Y, as well as RT, and also viral rRNA replicases have the conserved organization of the palm domain, but the folds of the fingers and thumb are quite variable. These features also appear in different order in the sequence. Although

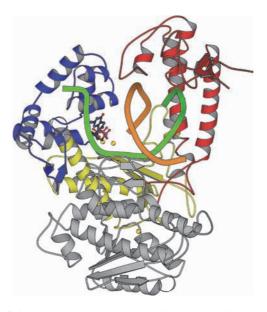


Fig. 9.18 ■ The structure of the T7 DNA polymerase (PDB: 1T7P). The structure can be described as a right hand. The palm is yellow, the fingers are blue and the thumb is red. The catalytically important magnesium ions are shown in yellow. The Exo domain (gray) is involved in proofreading. The DNA primer and template are colored orange and green, respectively.

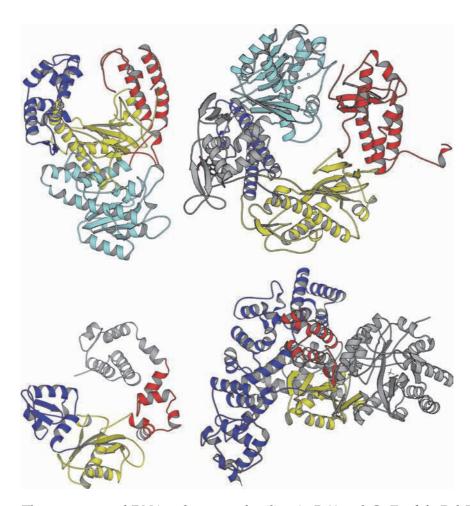


Fig. 9.19 ■ The structures of DNA polymerase families A, B X and C. Top left: Pol I, Klenow fragment (class A, PDB: 1KFS). Top right: DNA polymerase from phage RB69 (class B, PDB: 1IH7). Bottom left: Pol β (class X, PDB: 1PBX). Bottom right: Pol III α subunit (class C, PDB: 2HNH). The palm domain is yellow, the thumb domain is red and the fingers domain is blue. The top two polymerases have an exonuclease domain (pale turquoise). These domains are similar in structure but have strikingly different positions in the molecules. Some of the polymerases have additional domains, here shown in grey.

similar in general design, the polymerases of classes C and X have a different topology of the palm domain and a different order of the domains in the sequence (Figure 9.20).

9.2.4.3 The catalytic mechanism of DNA polymerases

The initial event in the DNA polymerization cycle is the binding of the DNA molecule to a primase that generates the enzyme-primer/template complex. Subsequently, DNA

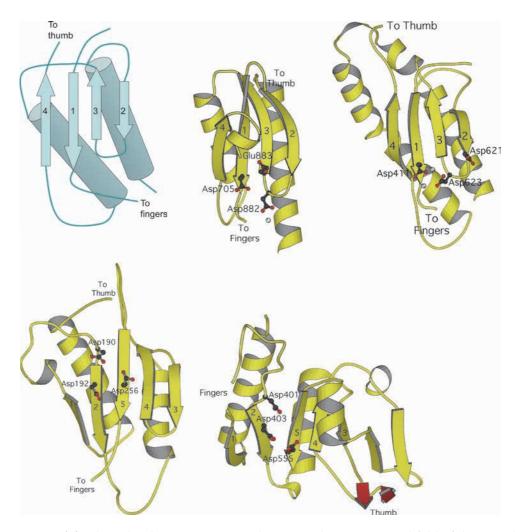


Fig. 9.20 ■ Top left: The palm domain in DNA polymerases has a conserved fold of the type double split β - α - β found in many nucleic acid binding proteins. Pol I (top middle) and RB69 polymerase (top right) have a different topology from that of Pol β (bottom left) and Pol III (bottom right). The three aspartate or glutamate residues involved in binding of the catalytic magnesium ions are shown.

polymerases (here the bacterial Klenow fragment) bind the rRNA/DNA hybrid. Then structural changes occur in the thumb sub-domain where the tip moves closer to the DNA molecule (Figure 9.21). The movement occurs in a helix-loop-helix motif that becomes ordered upon binding to DNA from a previously flexible structure. The palm domain interacts with the DNA minor groove and the 3'-end of the primer. The singlestranded DNA template (green) makes a sharp angle in its backbone. Thereby, it becomes located on the same part of the crevice between the fingers and thumb as the newly formed duplex DNA. Neither the template nor the duplex passes through the crevice or cylinder.

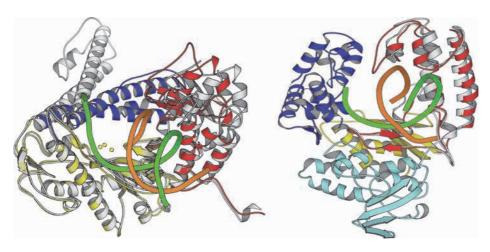


Fig. 9.21 ■ Examples of conformational changes in DNA polymerases upon binding of the template. Left: In the polymerase from bacteriophage RB69, the thumb (red) domain moves slightly towards the nucleic acid, while the fingers domain (blue) makes a large conformational change and moves residues used in catalysis towards the site of nucleotide addition (marked with the metal ions in yellow color). The apo enzyme is shown in grey. The N-terminal and exonuclease domains are removed for clarity (PDB: 1IG9 and 1IH7). Right: In the Klenow fragment of the Thermus aquaticus polymerase I, the thumb and its tip are bent towards the bound double-stranded DNA. The thumb domain of the apo structure is in grey. The other domains are similar in both structures and shown only for the DNA complex (PDB: 4KTQ and 5KTQ).

The next step is the nucleotide incorporation and starts when a dNTP binds to the enzyme-primer/template complex. Structures of Klenow fragments from different species are available with all four nucleotides bound to the N-terminal end of an α -helix of the fingers sub-domain called the O-helix. Initial recognition is done through the triphosphate moiety, which runs in parallel with the O-helix and interacts with positively charged arginines and lysines. A specific residue functions as a "sugar gate" and prevents the binding of rNTPs. The nucleotide base points towards the DNA-binding cleft. Once bound to the O-helix, the nucleotide is 10–15 Å away from the active site and a large conformational change or closure of the fingers domain delivers the nucleotide to the active center. The enzyme discriminates against incorrectly positioned or wrong nucleotides primarily through hydrogen bonding. Structures of dNTP complexes with several polymerases provide an insight into the rate-limiting step, related to the closure of the fingers subdomain. The efficiency in binding a correct or incorrect nucleotide varies a lot among polymerases. The selectivity of replicative polymerases is usually much greater and more important than for the repair polymerases.

In the active site, located in the palm domain, two magnesium ions (A and B) bound to conserved carboxylate residues stabilize the phosphates of the dNTP (Figure 9.22). The free hydroxyl group at the 3'-end of the growing strand activated by metal A attacks the

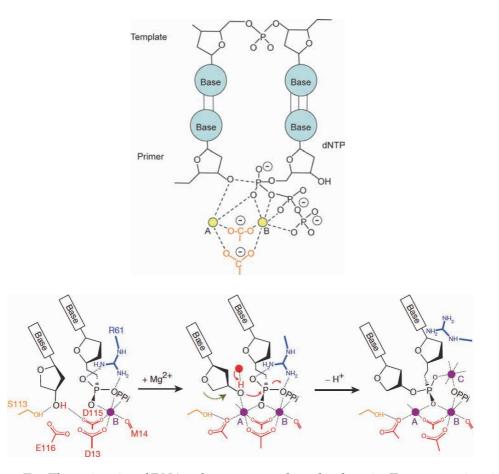


Fig. 9.22 ■ Top: The active site of DNA polymerases on the palm domain. Two magnesium ions A and B (yellow) participate in catalyzing the incorporation of new nucleotides according to the template. The magnesium ions are bound to two bridging aspartate residues (red). Metal A brings the 3'-OH and the bound nucleotide close to each other and activates the 3'-OH of the primer for attack on the α -phosphate of the dNTP molecule. Metal B binds to the α , β - and γ -phosphates of the dNTP and stabilizes the penta-coordinated transition state. (After Brautigam CA and Steitz TA. (1998) Curr Opin Struct Biol 8: 54-63). Bottom: A third magnesium ion participates transiently in the reaction by bridging between the α -and β -phosphates of the substrate nucleotide. (Reproduced with permission from Nakamura T, et al. (2012) Watching DNA polymerase η making a phosphodiester bond. Nature 487: 196–201. Copyright Macmillan Publisher Limited.)

phosphate of the dNTP, leading to the breakage of the bond to the β -and γ -phosphates. A new bond, between the α -phosphate and the free 3' hydroxyl group is formed.

In a time-resolved crystallographic investigation of pol η , a third magnesium ion transiently bridges between the α - and β -phosphate and the non-bridging oxygen of the

In eukaryotes, the primase (pol α) is composed of two primase subunits and two polymerase subunits. Its subunit pol 1 makes an rRNA strand of 10–12 nucleotides or about one turn of a helix and continues to add about 20 DNA nucleotides with its polymerase subunits. Subsequently, pol α falls off from the rRNA/DNA hybrid. Pol, δ or ϵ then continues to attach dNTPs to the primer until replication is completed. In the process, the rRNA primers are degraded and replaced by dNTPs, to make the new strands consist exclusively of DNA. Polymerases belonging to family B, including human pol α , may have a different initial recognition. Its nucleotide-binding site is located just at the template base in the active site and thus provides an immediate control of correct versus incorrect nucleotide binding through direct pairing.

The rate-limiting step of nucleotide incorporation is the conversion of the enzyme/primer/template: dNTP to an activated complex. This step is essential for the phosphoryl transfer reaction. The active site gets organized to allow the enzyme to proceed with the chemical reaction. This step (in family A enzymes) also plays a major role in discrimination between correct versus incorrect nucleotides. Quench flow studies on T7 polymerase and Klenow fragments show that the ratio of the rates for incorporation of correct versus incorrect nucleotides differ by 2000–5000 times, thus drastically slowing down the reaction in the case of an incorrect nucleotide incorporation.

In conclusion, when a correct nucleotide binds, the conformational change takes place and generates the proper configuration in the active center. Subsequently, the chemical reaction occurs (Figure 9.22). When an incorrect nucleotide binds, the conformational change takes place but the proper geometry of the active site is not reached. This slows down the chemical reaction significantly.

The phosphoryl transfer reaction is followed by a second conformational change, which releases the PPi product. Translocation of the primer/template DNA coincides with the opening of the fingers sub-domain. In the open state, the DNA molecule can move along the electrostatic tunnel forming the DNA-binding site, which is covered with positively charged residues. Thus, the newly formed base pair is rapidly moved in the opened complex to allow another cycle of incorporation to be initiated.

In humans, pol δ , the DNA polymerase handling the lagging strand, is composed of four subunits. Subunit Pol 3 is the catalytic subunit. Pol δ and pol ϵ are high fidelity polymerases (error rate 10^{-7}) with an exonuclease active site more than 40 Å away from the polymerase active site performing proofreading exonuclease activities. Here, the correctness of the five last incorporated nucleotides are proofread through specific hydrogen bonding in the minor groove, which only works for correct Watson–Crick base-pairing. Mutations in the exonuclease domain can frequently lead to development of cancers.

The catalytic subunit of pol ε (Pol 2) in humans is a large protein with two tandem polymerase/exonuclease parts. The palm domain of pol 2 is larger than other family B polymerases and has several extended sub-domains. Due to this interaction, Pol ϵ has high processivity even without a sliding clamp. The high fidelity of pol ε is maintained due to close interactions by parts of the protein around the base pair formed by the dNTP and the template nucleotide. These interactions are primarily van der Waals contacts. The exonuclease active site also contains two metal ions.

9.2.5 DNA Repair

DNA polymerases rarely introduce erroneous nucleotides. When it happens, DNA repair enzymes improve the error rate by a factor of 100. Furthermore, in any cell, the DNA structure changes all the time being constantly exposed to various physical, chemical and biological "attacks" introducing deviations from the normal double helical structures. Damage to the DNA can occur through endogenous or environmental factors or when the cell is exposed to radiation and mutagenic chemicals. DNA can also undergo spontaneous damage from the action of water, such as depurination/depyrimidation leaving abasic sites or deamination of cytosine and oxidation of guanosine. Malfunctioning enzymes can damage the DNA. The lesions may be single or double stranded breaks, loss of specific bases or chemical changes of bases. These changes represent a threat to the genetic constitution of the cell and are corrected by a range of enzymes, which constantly check the DNA to recognize any damage. Only a few of the relevant enzymes will be mentioned below.

During aerobic respiration, reduction of molecular oxygen to water generates partially reduced intermediates and by-products. Such radicals are potential electrophilic oxidants, which can attack various cellular components. One of the most vulnerable targets is DNA, and the oxidation of guanine bases into 7,8-dihydro-8-oxoguanine (8-oxoG) is very common (Figure 9.23). The modified base can make a Hoogsteen base pair (Section 5.3.3) with adenine, which will lead to a change or transversion of a G-C pair into a T-A pair (Figure 9.24). The generation and action of 8-oxoG is one of the main causes of mutations and spontaneous mammalian cell transformation, which can subsequently lead to development of various cancers. Therefore, the cell must defend itself against these changes.

9.2.5.1 DNA-repairing enzymes

One way to remove 8-oxoG from the pool of nucleotides is through the existence of a triphosphatase enzyme system, which specifically recognizes 8-oxo-dGTP and hydrolyzes it into 8-oxo-dGMP. However, if an 8-oxoG is found in the DNA it can be repaired

Fig. 9.23 ■ *Top*: The deamination of cytosine (*left*) generates uracil (*right*). *Bottom*: The oxidation of guanine (*left*) leads to 7,8-dihydro-8-oxoguanine (8-oxoG; *right*).

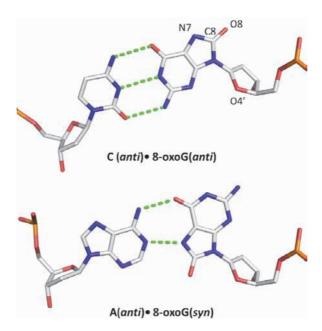


Fig. 9.24 ■ In the anti-conformation, 8-oxoG can make a normal base pair with a C. However, since the 8-oxo atom experiences a steric repulsion from the O4′ of the deoxyribose ring it prefers the syn-conformation where it would form a Hoogsteen base pair with an A (From Faucher F, et al., 2012).

by recognition, removal and replacement of the damaged unit. Here, several enzymes including mammalian 8-oxoG-DNA glycosylase (OGG1) recognize 8-oxoG opposite a cytosine in either the nuclear or the mitochondrial genome. Disruption of the enzyme gene results in accumulation of 8-oxoG in the genome and elevated rate of spontaneous mutagenesis. Humans express two alternatively spliced variants, of which one form is

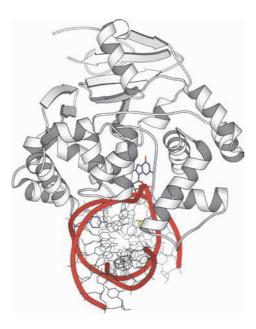


Fig. 9.25 ■ The structure of the DNA repair enzyme 8-oxoG-DNA glycosylase bound to DNA. The 8-oxo-G nucleotide is moved out of the double helix and into the active site of the enzyme. It and its base pair partner, a cytosine, are shown in blue (PDB: 1EBM).

transported into the nucleus and the other one into mitochondria. The structure of the human OGG1 bound to an 8-oxoG-containing DNA has provided answers to questions about the reaction mechanism (Figure 9.25).

To find a damaged base that is buried in the double helix, the enzyme scans the DNA by diffusing laterally along the minor groove until a damaged base is found. Somewhat surprisingly, the 8-oxo group is not identified by most forms of OGG1, but rather the protonated N7. The 8-oxoG is flipped out of the double helical structure and into the glycosylase active center with numerous interactions. The vacated position in the dsDNA is suitably filled by parts of the protein. A conserved aspartate residue functions as the catalytic residue in the reaction. This is the first step in the base excision repair (BER) pathway that repairs single base errors (lesions) in DNA. In the next step, an endonuclease cleaves an abasic site of the DNA on the 5' side of the site. The lyase domain of pol β (family X) then removes the damaged nucleotide and the polymerase domain of pol β inserts a nucleotide with proper Watson-Crick base pairing to the opposite strand. Finally, a ligase restores the original structure of the DNA.

Several other mechanisms are needed for the repair of damaged DNA. For example, a damaged nucleotide (and not only the base) can be recognized and removed from the DNA backbone. This is called nucleotide excision repair (NER). A different set of enzymes performs this repair and the recognition is based on a different principle: a distortion in the double helix structure is recognized rather than a single modified base. Yet another repair mechanism is based on homologous recombination.

9.3 **Telomerases**

Linear chromosomes are replicated differently from circular ones (found in most bacteria), since the ends not only have to be replicated but also protected against degradation. DNA polymerase can only synthesize a new strand of DNA as it moves along the template strand in the 3′ -> 5′ direction (Figure 9.26). This works fine in the forward direction, but the lagging strand is replicated discontinuously. This continues until, close to the end of the DNA, the template is too short to form Okazaki fragments. In this way, the 5'-end of each newly synthesized strand cannot be completed, which sometimes results in chromosome shortening. After a few generations, genes could be truncated or lost leading to replicative senescence, which can block cell division.

Eukaryotic chromosomes have several DNA elements that are not genes and not directly involved in regulating the gene expression. These elements include origins of replication, centromeres that are involved in the movement of chromosomes into daughter cells, and telomeres that replicate and protect the chromosome ends.

Telomeres are located at both ends of a linear chromosome. Eukaryotic telomeres consist of short sequences, only a few bases in length, repeated hundreds to several thousand times and terminating with single-stranded 3' overhangs. A common repeat is the 6 bases TTAGGG, which could form quadruplex structures (Section 5.3.4). Several proteins bound at the chromosome ends can form unique structures to distinguish the chromosome ends

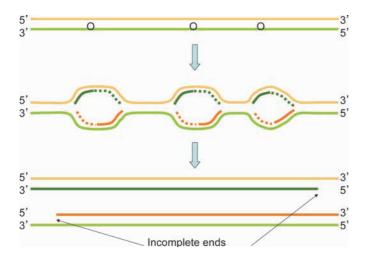


Fig. 9.26 ■ Replication of a piece of DNA with three origins of replication (O). Starting from an origin of replication the leading strand will be copied continuously. The lagging strand has to be synthesized as fragments (See also Figure 9.6). This causes a problem close to the ends and the 5' ends of the new DNA strands (darker color) could stay incomplete. The telomerase will assist to complete the ends of the newly synthesized strands.

from sites of chromosome breakage. Thereby, the telomeres are protected from recombination and degradation, which could severely impair the chromosome structure. Telomeres also act as specialized origins of replication that allow the cell to replicate the chromosome end. The unusual DNA polymerases involved in the telomere replication are called telomerases, found primarily in germ cells and certain stem cells.

Human telomerase is a ribonucleoprotein complex with a size of 670 kDa. Its main function is the reverse transcriptase activity (TERT), converting rRNA-encoded information into DNA. The rRNA component of the enzyme (TER) is used as a template for reverse transcription. The resulting DNA is added to the single stranded overhangs, which could not be replicated during the normal replication cycle. TER thus represents a "memory" element with complementary sequence to the repeats of the single stranded overhangs at the chromosome ends. Thus, TER can again and again rescue the chromosomal ends. The 3'-OH group of the terminal nucleotide at the telomere end is the primer for reverse transcription. TER in combination with TERT contributes to proper nucleotide addition and processivity in the addition of telomere repeats, to oligomerization and to nuclear localization (Figure 9.27). Apart from TERT and TER, several accessory proteins are involved in telomerase assembly, accumulation, recruitment to the telomere and other functions.

The size of TER varies greatly among different organisms. It is relatively short in ciliates, approximately 150 bases, around 450 bases in vertebrates and can be over 1300 in yeast. However, all known TERs contain the main elements: (i) the pseudoknot-template part and (ii) a stem-loop element. Elements of the 3D structure of TER are known and confirm the secondary structures predicted.

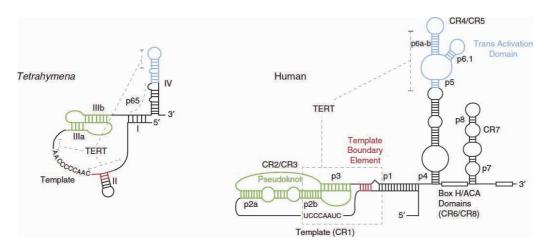


Fig. 9.27 ■ The secondary structure of telomerase rRNA (TER) from a ciliate, Tetrahymena thermophila and H. sapiens. The pseudoknot is shown in green and the stem-loop in blue (Sandin S, Rhodeso, 2014).

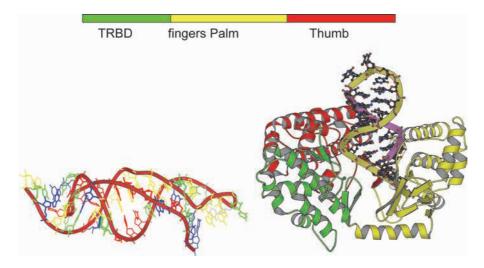


Fig. 9.28 ■ Top: The domain organization of TERT from Tribolium castaneum. Bottom left: The pseudoknot from human TER. The structure contains nucleotides from stems p3 and p2b (see Figure 9.27). The nucleotides are colored blue (guanine), green (cytosine), yellow (adenine) and red (uracil). Two nucleotides that were used in the construction of the rRNA molecule are shown in grey. The central part of the molecule has a triple-helical structure (PDB: 1YMO). Bottom right: The structure of TERT from T. castaneum in complex with a fragment of the TER rRNA (olive) and a matching piece of the DNA (purple PDB: 3KYL).

9.3.1 The Telomerase Holoenzyme

TERTs from various organisms have about 1100 amino acid residues with several motifs in common, but also contain distinct species-specific parts. The structure of the reverse transcriptase domain (RT) is similar to DNA polymerases but particularly similar to viral reverse transcriptases with several universally conserved elements. The domains are called, from the N-terminus: TEN, a variable linker, TRBD, the reverse transcriptase part including fingers and palm and finally CTE (which includes the thumb; Figure 9.28, top). The fingers domain is involved in nucleotide and rRNA binding, the palm domain contains the active site, the TEN domain and the thumb bind the single-stranded DNA of the telomere and TRBD binds to the rRNA pseudoknot (Figure 9.28). Some species lack the TEN domain. TERT forms a tunnel structure with the catalytic residues on the inside. The tunnel is large enough to enclose about 8 base pairs of rRNA/DNA.

TER and TERT alone can reconstitute the telomerase enzymatic activity. Human telomerase seems to function as a dimer. In the cell, however, a larger complex containing additional proteins operates.

9.3.2 Enzymatic Properties

Telomerases act like other reverse transcriptases. A unique feature, however, is that they use a template, which is part of the TER rRNA molecule. They are also polymerases and

can add multiple complements of the template through several cycles of extension and translocation reactions. Firstly, the chromosome end is recognized by the telomerase holoenzyme and the DNA 3'-end makes a base-paired hybrid with the rRNA template. Secondly, the template directs addition of nucleotides to the 3'-end of DNA until the 5'-end of the template is reached. A unique step follows; the enzyme undergoes translocation and moves to the new 3'-end of the growing DNA strand. Another round of copying of the rRNA template takes place. Multiple translocations and nucleotide additions result in formation of several repeats.

Telomerase chemistry of nucleotide transfer is, like in other DNA polymerases, based on a two-metal mechanism. Like for other polymerases, the catalytic metal is magnesium, bound to conserved Asp and Glu residues.

Recombination 9.4

Different DNA molecules can be recombined. The primary case is during meiosis where homologous chromosomes are paired and genetic information is exchanged. This generates new genetic information that can be passed on to new generations and is one of the main processes driving evolution. A pair of homologous chromosomes, one originating from the mother and another from the father, combines in the eukaryote cell prior to the first division. During the pairing genetic exchange occurs (also called crossing over) and represents the main event of homologous recombination (HR). HR is a process catalyzed by enzymes specifically synthesized and regulated for this particular purpose. In addition, recombination also allows cells to (i) retrieve sequences lost through DNA damage, (ii) restart stalled replication and (iii) to regulate the expression of some genes. The central methods of molecular genetics, which are used to delete specific genes and create directed mutants, are based on homologous expression, for example, generation of "knock-out" mice. The RecA protein and related proteins is central in HR.

9.4.1 RecA and Related Recombinases

DNA pairing and strand-exchange are the main activities during HR. In bacteria, RecA is a key enzyme involved in the search for sequence matches between two DNA molecules leading to base pairing between homologous parts of these two molecules. RecA belongs to a family called strand-exchange proteins or recombinases, and similar proteins are found in almost all organisms. In eukaryotes, the protein Rad51 performs the same function.

RecA has a main domain with what is called the RecA fold (Figure 8.14). It binds ATP and is related to other P-loop-containing proteins, like helicases and F_1 ATPases (Section 8.3). In RecA and Rad51, a glutamate residue participates in ATP hydrolysis.

RecA and Rad51 can exist as helical inactive filaments in the absence of DNA. The monomers are arranged in a six-fold helical structure. The DNA-binding loops of the protein are located on the inner side of the helix. The pitch of the helix is around 75 Å.

ATP and ssDNA can bind to interfaces between monomers of RecA in a cooperative manner. The helical pitch here will then change to around 95 Å. Large filaments are formed, which can interact with other strands of DNA. The size is variable and over 100 RecA monomers can participate in an active complex. The N- and C-terminal domains are involved in both DNA binding and stabilization of the oligomeric structure.

9.4.2 Recombinases in Action

The action of recombinases like RecA starts after a double stranded break, where the nucleotides on the 3′ side are being digested away, leaving a single strand of DNA. The ssDNA rapidly binds ssDNA-binding proteins (in *E. coli* called SSB), preventing the ssDNA from being further degraded as well as melting secondary structure. SSB is subsequently replaced by recombinases (RecA) to form the presynaptic filaments with the aid of mediator proteins (Figure 9.29).

In an elegant set of experiments the structures of RecA filaments with both ssDNA and dsDNA have been revealed. Several RecA molecules were expressed as a single polypeptide, where the separate RecA molecules were connected by flexible linkers. The protein was in complex with an ATP analog (ADP-AlF₄) and oligomers of ssDNA or dsDNA. In all cases, the filament axis is essentially straight with the Watson–Crick edge

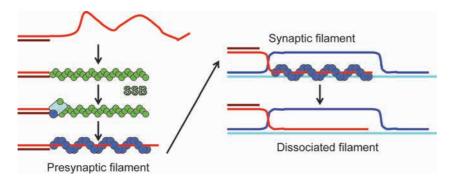


Fig. 9.29 ■ The process starts from ssDNA where SSB (green) molecules bind, protect and melt secondary structures. SSB is subsequently replaced by RecA (blue) or corresponding proteins to form the presynaptic filaments. The presynaptic filaments can recruit dsDNA and search for homologous DNA that can be recombined with the ssDNA (From Liu J, *et al.* 2011).

of the bases located close to this axis. The helical repeat was 6.16 RecA molecules per turn with a helical pitch of about 94 Å. The ATP analog binds to the Walker motifs between RecA molecules. The AlF₄ or γ-phosphate stabilizes the active RecA-RecA interface that can bind DNA. ATP hydrolysis and release of the inorganic phosphate will result in a different conformation of the RecA filament releasing the bound DNA. In both inactive and active filaments, the core of RecA interacts with the N-terminal domain of the next RecA molecule, but in the active filament the core has undergone a significant rotation and translation.

In the case of the ssDNA, the structure of the presynaptic filament, nucleotide triplets are bound to each RecA molecule to form separate units (Figure 9.30). The average helical parameters of the DNA in the filament are 18.5 nucleotides per turn with a helical rise of 5.08 Å per base or base pair. The triplets have a B-DNA-like conformation with a right-handed twist and an axial rise of 4.2 Å, but between triplets there is an axial rise of 7.8 Å and a left-handed twist making the DNA a more open structure. The backbone of the DNA interacts with the protein and the Watson-Crick edge exposed to the solvent.



Fig. 9.30 ■ Left: The crystal structure of a filament of six RecA molecules in different colors in complex with six ADP-AIF₄ molecules and (dT)₁₈ (PDB: 3CMU). The ssDNA (red) has an open conformation and follows the helical axis. Arrows indicate the position of ATP in the RecA molecules.

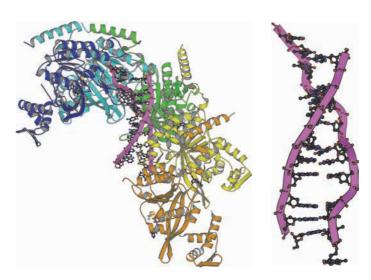


Fig. 9.31 ■ *Left*: The structure of five RecA molecules in complex with ADP-AlF₄ and $(dT)_{15}$ $(dA)_{12}$ (PDB: 3CMX). The dsDNA follows the filament axis. The structure is closely similar to the case with ssDNA. *Right*: The structure of the dsDNA with triplets of base pairs stacked like in B-DNA.

RecA with bound dsDNA is easily superimposable on RecA with ssDNA (Figure 9.31). Again, the DNA is arranged around the filament axis. The triplet arrangement of stacked base pairs is retained and the open structure remains. The complementary strand has the antiparallel orientation and the bases interact through Watson–Crick hydrogen bonds. The primary strand of DNA has essentially the same conformation as in the structure with ssDNA and the base pairs of the triplet closely resemble B-DNA (Figure 9.31). The helical rise from one triplet to the next is 8.4 Å with an underwound twist. The complementary strand has limited contact with RecA. Thus, the heteroduplex formation is highly dependent on the correct base pairing and non-Watson–Crick base pairs would fit poorly.

What remains to be established is how the donor dsDNA binds to the presynaptic filament. Two basic residues, Arg243 and Lys 245, are implicated in this activity and are found about 25 Å away from the filament axis and their distance to the same pair on the next RecA molecule is about 28 Å (Figure 9.32). The surface of the filament has an electropositive groove into which the dsDNA could bind. Modeling of the DNA into the groove with interactions with these basic residues suggests that consecutive base pairs cannot stack properly. Furthermore, since this secondary site has lower affinity for dsDNA than ssDNA it suggests that the dsDNA in this site is destabilized and that the Watson–Crick base pairing with the primary strand can be sampled. When a homologous DNA is found, cleavage, ligation and synthesis of complementary DNA to single-stranded bits will occur with the appropriate enzymes of the species in question.

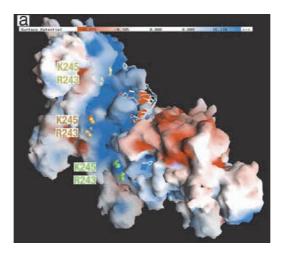


Fig. 9.32 ■ The structure of five RecA molecules with bound dsDNA showing a set of exposed basic residues separate from the bound dsDNA, which may be the binding site for the secondary or incoming dsDNA. The electrostatic properties of the surface of the RecA pentamer shows the positively charged groove that is likely to bind the incoming dsDNA. (Reproduced with permission from Chen Z, et al. (2008) Mechanism of homologous recombination from the RecA-ssDNA/ dsDNA structures. Nature 453: 489–494. Copyright Nature Publishing group.)

Further Reading

Original Article

Arias-Palomo E, O'Shea VL, Hood IV, Berger JM. (2013) The bacterial DnaC helicase loader is a DnaB ring breaker. *Cell* **153**: 438–448.

Beese LS, Derbyshire V, Steitz TA. (1993) Structure of DNA polymerase I Klenow fragment bound to duplex DNA. Science 260: 352-355.

Chen Z, Yang H, Pavletich NP. (2008) Mechanism of homologous recombination from the RecAssDNA structure. Nature 453: 489-494.

Costa A, Renault L, Swuec P, et al. (2014) DNA binding polarity, dimerization, and ATPase ring remodeling in the CMG helicase of the eukaryotic replisome. *eLife* **3**: e03273, 1–17.

Franklin MC, Wang J, Steitz TA. (2001) Structure of the replicating complex of a pol alpha family. Cell 105(5): 657-667.

Itsathitphaisarn O, Wing RA, Eliason WK, et al. (2012) The hexameric helicase DnaB adopts a nonplanar conformation during translocation. Cell 151: 267–277.

Kelch BA, Makino DL, O'Donnell M, Kuriyan J. (2011) How a DNA polymerase clamp loader opens a sliding clamp. Science 334: 1675–1680.

Kitani K, Kim S-Y, Hakoshima T. (2010) Structural basis for DNA strand separation by the unconventional winged-helix domain of RecQ helicase WRN. Structure 18: 177–187.

- Lamers MH, Georgescu RE, Lee SG, et al. (2006) Crystal structure of the catalytic alpha subunit of E. coli replicative DNA polymerase III. Cell 126: 881–892.
- Liu B, Eliason WK, Steitz TA. (2013) Structure of a helicase-helicase loader complex reveals insights into the mechanism of bacterial primosome assembly. Nat Commun 4: 24951-24958.
- Liu J, Ehmsen KT, Heyer W-D, Morrical SW (2011) Presynaptic filament dynamics in homologous recombination and DNA repair. Crit Rev Biochem Mol Biol 46: 240-270.
- Mitchell M, Gillis A, Futahashi M, et al. (2010) Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. Nat Struct Mol Biol 17: 513-518.
- Nakamura T, Zhao Y, Yamagata Y, et al. (2012) Watching DNA polymerase H make a phosphpodiester bond. Nature 487: 196-201.
- Parikh SS, Walcher, G, Jones, GD, et al. (2000) Uracil-DNA glycosylase-DNA substrate and product structures: Conformational strain promotes catalytic efficiency by coupled stereoelectronic effects. Proc Natl Acad Sci USA 97: 5083-5088.
- Song F, Chen P, Sun D, et al. (2014) Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosome units. Science 344, 376-380.
- Story RM, Weber IT, Steitz TA. (1992) The structure of the E. coli rec A protein monomer and polymer. Nature 355: 318-324.
- Xia S, Koningsberg H. (2014) RB69 DNA polymerase structure, kinetics and fidelity. Biochemistry **53**: 2752–2767.

Reviews

- Chang C-C, Wang Y-R, Chen S-F, et al. (2013) New insight into DNA-binding by type IIA topoisomerases. Curr Opin Struct Biol 23: 125-133.
- Cox MM. (2007) Monitoring along with the bacterial RecA protein. Nat Rev Cell Mol Biol 8: 127-137.
- Gubaev A, Klostermeier D. (2014) The mechanism of negative DNA supercoiling: A cascade of DNA-induced conformational changes prepares gyrase for strand passage. DNA Repair 16: 23-34.
- Gurard-Levin ZA, Quivy J-P, Almouzni G. (2014) Histone chaperones: Assisting histone traffic and nucleosome dynamics. Ann Rev Biochem 83: 487-517.
- Johnson A, O'Donnell M. (2005) Cellular DNA replicases: Components and dynamics at the replication fork. Ann Rev Biochem 74: 283-315.
- Mason M, Schuller A, Skordalakes E (2010) Telomerase structure and function. Curr Opin Struct Biol **21**: 92–100.
- Onesti S, MacNeill SA (2013) Structure and evolutionary origins of the CMG complex. Chromosoma **122**: 47–53.
- Sandin S, Rhodes D (2014) Telomerase structure. Curr Opin Struct Biol 25: 104–110.
- Singleton MR, Dillingham MS, Wigley DB. (2007) Structure and mechanism of helicases and nucleic acid translocases. Ann Rev Biochem 76: 23-50.

10

Transcription

Transcription is the cellular process where rRNA is synthesized by DNA-dependent rRNA polymerases from the four ribonucleoside triphosphates using DNA as a template. In eukaryotes, the regulation of transcription is central for cell differentiation and organism development. Thus, the complex regulatory system is equally important as the rRNA polymerase. The direction of the synthesis is from 5′ to 3′. The rRNA is a complementary copy of the template strand of DNA through Watson–Crick base-pairing. Furthermore, it is an exact copy of the non-template DNA strand except that the thymine base of the DNA is replaced by uracil in rRNA. There are also RNA-dependent rRNA polymerases, for example, in viruses with rRNA genomes. The rRNA molecules synthesized are either used as tools in the cell such as tRNA, ribosomal rRNA or molecules containing information such as messenger rRNA used for translation, or RNAi and microRNA controlling gene expression.

As all polymerizing processes, transcription can be considered as three separate steps, initiation, elongation and termination. In order to copy the DNA, the transcription start site (TSS), has to be identified. The rRNA polymerase is itself unable to initiate transcription efficiently and specifically. This process is therefore aided by a range of proteins that act as activators or repressors for the initiation. However, viral transcription is usually performed by single subunit rRNA polymerases without the help of additional protein factors.

10.1 Control Elements in DNA

Transcription is a central process in all cells. The synthesis of all molecules in the cell depends on the transcription of appropriate molecules. Depending on the state of the cell and its environment, different needs have to be met. Therefore, transcription needs to be

10.1.1 Bacterial Promoters

Genes have specific sequences both before and after the coding region by which transcription can be controlled. The primary controlling elements are the promoters, which are segments of DNA located on the 5′-side of the coding region. In bacteria, the promoters are conserved sequences upstream of the transcriptional start site to which the rRNA polymerase binds. In *E. coli*, two of them are centered at about –10 bp (Pribnow box, with a consensus sequence TATAAT) and –35 bp (consensus sequence TTGACA), with a third site further upstream. Activators and repressors can also bind to the promoter region of the controlled gene and stimulate or prevent, respectively, the binding of the rRNA polymerase to transcribe the gene.

10.1.2 Promoters and Regulatory Elements in Eukaryotes

In eukaryotes, there is no universal motif that defines a transcription start site (TSS). Different type of genes, related to specific tissues or developmental stages, can have different type of promoters.

The initiation of transcription can be done in a focused manner within a narrow region or in a dispersed manner over a number of weak start sites in a broad region of upto 100 nucleotides. One focused DNA element is the TATA box, found in only 10–20% of all genes. The first T is located about 30 nucleotides before the transcriptional start site (TSS). The TATA-box containing promoters have low GC content and are single or narrow TSS promoters. Important to regulate transcription of genes are the gene-specific transcription factors that activate or repress the initiation of the process by binding to regulatory elements that can be found upto several kbases from the gene (Figure 10.1).

10.1.3 Control of Transcription through DNA Packaging

DNA needs to be packaged and protected against degradation, while it also needs to be accessible for transcription. Eukaryotic DNA is located in the nucleus and compacted into

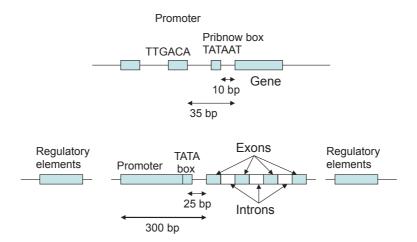


Fig. 10.1 ■ *Top*: The outline of a bacterial gene. The promoter contains some general elements and a specific region identifying the gene. *Bottom*: The outline of a eukaryotic gene. The transcribed region can be composed of both exons and introns. The introns are removed after transcription. Different elements before and after the transcribed part are involved in the regulation of the gene.

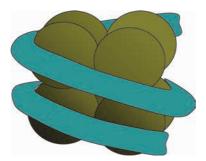


Fig. 10.2 ■ A simplified representation of a nucleosome with the histone octamer inside (green) and about two turns of DNA (blue) wrapped around it in a left-handed manner.

chromatin, where the DNA is wrapped around the histone octamers to form nucleosomes (Figure 10.2), which in turn are organized into higher-order structures (see Section 9.1). In order to be transcribed, the promoter of a gene has to be accessible. Obviously, transcription is to a large extent regulated by the structural organization of the genome.

10.1.3.1 Chromatin remodeling

The nucleosomes are distributed along the DNA in a non-random fashion. Short linker-DNA segments, with lengths of 20–50 bp, connect the nucleosomes. Nucleosome mapping show that several factors influence the position of the nucleosomes, both the sequence of nucleotides and the binding of numerous proteins (Figure 10.3). Genes can be silenced by high nucleosome occupancy, whereas frequently transcribed genes can be

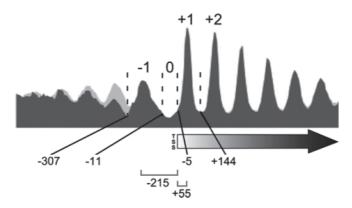


Fig. 10.3 ■ Mapping of nucleosome positions in yeast in relation to the TSS (position 0). The most sharply defined nucleosome position is +1. (Reproduced from Richmond T. (2012) Nucleosome recognition and spacing by chromatin remodeling factor ISW1a. Biochem Soc Trans 40: 347-350. Copyright (2012) Portland).

devoid of nucleosomes. Essential genes for growth, such as the genes for ribosomal proteins, often have fewer nucleosomes. Regulatory elements like promoters can be found in nucleosome-free regions (NFR) or nucleosome-depleted regions (NDR). In yeast, there is normally a 140-bp NFR at the promoter site between nucleosomes –1 and +1. Nucleosomes flanking promoters and enhancers are usually well positioned, but the accuracy in nucleosome position is less precise with increasing distance from the regulatory elements (Figure 10.3).

The DNA sequence can influence the placement of nucleosomes. A high GC content increases nucleosome occupancy, while a high AT content leads to nucleosome depletion. Thus, growth genes in yeast normally have AT-rich promoters. On the other hand, short bendable AT dinucleotides face inwards in the nucleosome, while the stiff GC dinucleotides face outwards. Thus, a 10-bp AT dinucleotide periodicity can increase the nucleosome occupancy.

While the sequence dependence of nucleosome positioning is called cis-acting factors, the trans-acting factors are: ATP-dependent chromatin remodelers, TFs and rRNA polymerase. The remodelers can induce nucleosome sliding, partial or complete nucleosome eviction, or exchange histone dimers from the nucleosome (Figure 10.4). There are numerous remodeling ATPases that can be grouped into four major families. The basic arrangement of the ISWI type is shown in Figure 10.5.

The nucleosomes move on the DNA through sliding, as has been seen by single molecule fluorescence resonance energy transfer (FRET; Figure 10.6).

Histones have variants, which affect the accessibility of the genes, and are positioned in specific ways. A well-known variant is histone H2A.Z. A modified nucleosome may be homotypic with two H2A.Z histones or heterotypic with an H2A/H2A.Z arrangement. The homotypic nucleosomes have higher stability and less access to the DNA, while the heterotypic arrangement is less stable. While the homotypic version frequently is placed

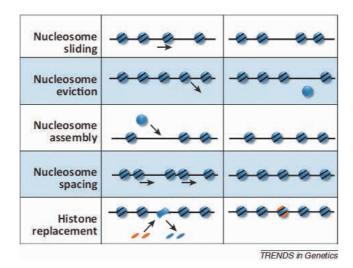


Fig. 10.4 ■ Chromatin remodeling ATPases and other proteins can affect the positions of nucleosomes by a number of mechanisms. (Reproduced with permission from Petty E & Pillus L (2013.) Balancing chromatin remodeling and histone modifications in transcription. *Trends Z Genetics Z* **29**: 621–629. Copyright Elsevier.)

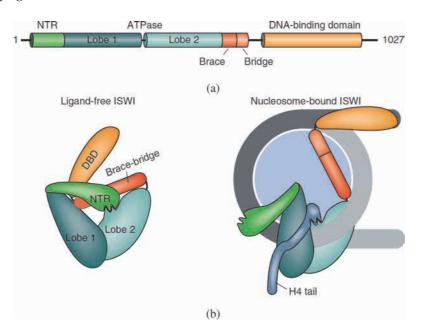


Fig. 10.5 ■ (a) The domain organization of the ISWI type of remodelers. The N-terminal region (NTR) is followed by two lobes of the ATPase. Before the DNA-binding domain (DBD) there are two smaller elements called Brace and Bridge. (b) The enzyme undergoes a conformational change in going from the idle enzyme to when it binds to a nucleosome. While DNA-binding domain binds at SHL-7, the ATPase lobes bind to the superhelical location −2 (SHL-2). (Reproduced with permission from Mueller-Planitz F *et al.* (2014) Nucleosome sliding mechanisms: New twists in a looped history. *Nat Struct Mol Biol* 20: 1026–1032. Copyright (2014) Nature Publishing Group.)

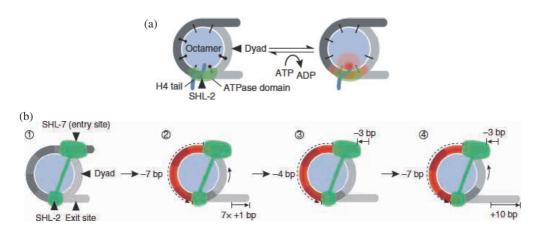


Fig. 10.6 ■ *Top*: The binding of the ATPase domain of ISWI to a nucleosome occurs at SHL-2 and distorts the contacts between the DNA and the histones. *Bottom*: The complete enzyme interacts not only by the ATPase, but the DNA-binding domain interacts at the SHL-7 on the opposite side and induces a ATPase-dependent sliding of the DNA on the histone core. (Reproduced with permission from Mueller-Planitz F *et al.* (2014) Nucleosome sliding mechanisms: New twists in a looped history. *Nat Struct Mol Biol* **20**: 1026–1032. Copyright (2014) Nature Publishing Group.)

as nucleosome +1 after TSS, the unstable heterotypic arrangement may be found even at the TSS site of an active gene. The H2A.Z histone is incorporated in a stepwise fashion.

10.1.3.2 Histone modifications and epigenetics

The N-termini of histones extend from the core of the protein. They are flexible in single nucleosomes and can be modified through methylation, acetylation, phosphorylation, ubiquitination and SUMOylation (see Textbox).

Enzymes are adding (writers) or removing (erasers) the histone modifications, which can be identified by the gene-specific protein modules (readers) that recruit proteins that can influence the packing of the nucleosomes and play important regulatory roles during transcription. The protein modules distinguish the pattern of modifications and in turn recruit other proteins, which can regulate the transcription of certain genes. As part of the recognition, these modules bind parts of the histone tails through β -strand interactions.

Reading of methylated lysines is complex since one (me1), two (me2) or three methyl groups (me3) on a single ϵ -amino group of a lysine have to be distinguished from each other and also from non-methylated lysine (me0). A methylated lysine has a reduced number of hydrogen bond possibilities. In particular, me3 cannot engage in any hydrogen bonding but remains positively charged. Certain protein modules can specifically identify the different levels of methylation on lysines. To these belong the chromo and tudor domains. The binding pockets of these modules are built of several aromatic side chains. The clouds of π -electrons over the aromatic rings neutralize the lysine charge while

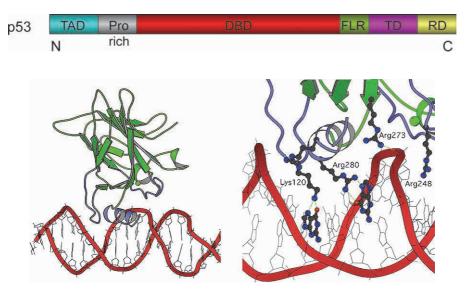


Fig. 10.17 ■ *Top*: The domain organization of p53. The transcription activation domain (TAD) can be extensively phosphorylated. The DBD domain is the DNA-binding core domain. FLR is a flexible linker region next to the tetramerization domain (TD). RD is the regulatory domain. The two last domains can be acetylated. The structure model of p53 was generated by homology modeling and molecular dynamics simulation (From Saha T, *et al.*, 2015). *Bottom left*: The structure of the DBD domain of p53 interacting with DNA. While a helix and a loop are fitted into the major groove, a different loop interacts with an adjacent minor groove. *Bottom right*: A detailed view showing some of the side chains interacting with the DNA. Arg280 forms the most critical contact with a guanine (PDB: 1TUP). Arginine 273, which interacts with an oxygen atom in a phosphate and Arg 248, which is sandwiched in the minor groove, are both frequently mutated in tumors.

p53 binds to DNA regions that typically have four copies of the consensus sequence. It is possible that the DBDs of a tetramer of p53 bind to each of these four adjacent copies.

10.2.7 Binding Specificity

The specificity in binding proteins to DNA depends on hydrogen bonds between side chains and bases, as well as on interactions that depend on the backbone conformation of the DNA. Since each base pair in the DNA presents a unique pattern of hydrogen bond donors and acceptors in the major and minor grooves (Figure 10.18), they can be recognized by side chains like arginines and glutamines, which also have distinct patterns of hydrogen bond donors and acceptors.

Although most TFs use helices binding in the major groove of the DNA for specific recognition, there are examples of proteins that bind differently. One example is the Arc repressor from bacteriophage P22. This protein is a tetramer, and a pair of monomers forms a two-stranded antiparallel sheet that binds in the major groove of DNA (Figure 10.19).

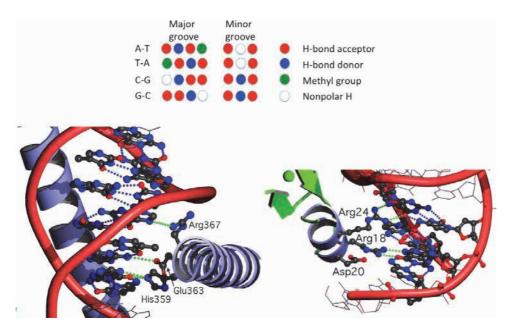


Fig. 10.18 ■ *Top*: The exposed groups in the major and minor grooves of dsDNA. *Bottom left*: The interaction of Myc with DNA (see also Figure 10.11). The heterodimer binds specifically to a hexanucleotide segment CACGTG. Three residues in the helix make specific interactions with three base pairs forming half of the recognized palindromic sequence, while the same residues in the other monomer binds similarly to the other half of the segment (not shown). The arginine is especially important in recognizing the CG base pair at the third position in the sequence. Hydrogen bonds between protein and DNA are green; hydrogen bonds in the base pairs are blue. Further interactions with the DNA backbone stabilize the contact (PDB: 1NKP). *Bottom right*: The first zinc finger of Zif 268 interacting with DNA (see also Figure 10.15). The two arginine residues interact with guanine bases. Asp20 interacts with Arg18, keeping it in position.

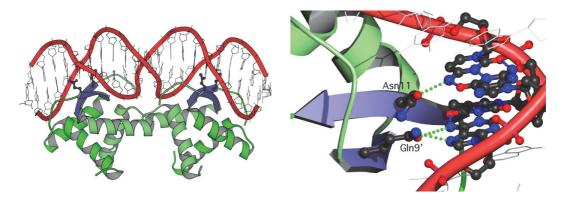


Fig. 10.19 ■ *Top*: The arc repressor tetramer. *Bottom*: A detail of the interaction showing the recognition of an adenine base by a glutamine side chain. An asparagine binds to the amino group of a cytosine (PDB: 1BDT).

10.3 **Bacterial Transcription**

In bacteria, there is one rRNA polymerase while there are three in eukaryotes. The bacterial enzyme is composed of five subunits of four kinds ($\alpha_2\beta\beta'\omega$) forming the core of rRNA polymerase (RNAP) and a fifth subunit, σ or sigma, which participates in recognizing the promoter and initiating transcription. RNAP including σ is called the holoenzyme. The molecular mass of the complex is around 400-450 kDa. Bacterial activators and repressors (Section 10.2) interact directly with the rRNA polymerase. Transcription in general goes through three main steps: initiation, elongation and termination.

First, the double-stranded DNA needs to be unwound for the interaction with the rRNA polymerase (Figure 10.20). The unwound region is called the "transcription bubble" and contains about 15 base pairs (bp). The transcribed rRNA forms a hybrid helix

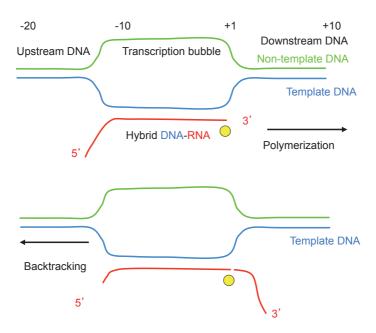


Fig. 10.20 ■ Transcription of DNA to RNA requires that a region of the duplex DNA is melted or unwound. This region to which the RNA polymerase is bound is called the transcription bubble. During transcription, a DNA-RNA hybrid of 8 to 9 base pairs is formed in the transcription bubble. The downstream DNA is not yet transcribed, while the upstream DNA has regained the duplex form after transcription. The RNA polymerase can move forward as it does during RNA polymerization, but backtracking is also possible during which incorrectly incorporated nucleotides can be excised. The yellow circle marks the active site where nucleotides are added to the RNA chain. The numbering is in relation to the active site of the polymerase rather than to the start of the transcribed gene.

with the template strand of DNA, with a length of about eight or nine base pairs. Bacterial as well as eukaryotic rRNA polymerases are capable of both forward and backwards movements. The binding of nucleoside triphosphates stimulates the forward movement, while damaged DNA can cause backtracking.

During initiation of transcription, the holoenzyme binds to two hexameric sequences in the promoter region of the DNA at positions –35 and –10 in relation to the transcription start site (+1), with the aid of the σ initiation factor (see Section 10.2.2). Upon binding, the holoenzyme unwinds the double-stranded DNA between positions –12 and +2, leading to an open promoter complex from which transcription can be initiated. When a DNA-RNA hybrid of nine base pairs has been produced, the transcriptional initiation is over and the elongation phase begins. The elongation complex has high stability, partly due to a bound factor called NusG, and can synthesize rRNA chains thousands of nucleotides in length. In eukaryotes, the role of NusG is dealt with by a conserved factor called Spt5, which is part of a heterodimer, Spt4/5. These proteins lock the DNA to the rRNA polymerase.

10.3.1 Structure of the Bacterial rRNA Polymerase

The structure of RNAP is like a crab-claw with a length of about 150 Å and about 110 Å in both other dimensions (Figure 10.21). The structure can be described as having four rigid modules: core, shelf (the lower jaw), clamp and jaw-lobe (the upper jaw), which can move with regard to each other. The core, with the active site, and shelf modules form the base from which the clamp and jaw-lobe form the pincers of the claw. The core is formed by parts of the β , β' and the two α subunits. The other three modules surround the DNA-binding cleft. The incoming double-stranded DNA goes into the cleft, becomes unwound or melted, and is transcribed into a strand of rRNA. The exiting DNA comes out in a perpendicular direction in relation to the incoming DNA.

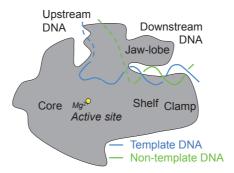


Fig. 10.21 ■ A schematic illustration of the bacterial RNA polymerase with DNA and the four main modules.

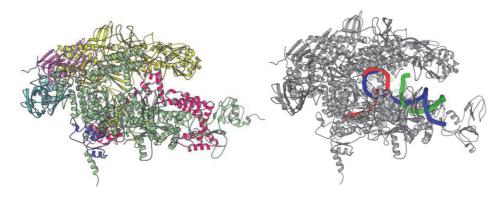


Fig. 10.22 ■ Left: The structure of the holoenzyme RNA polymerase from T. thermophilus (PDB: 1IW7). The β (yellow) and β' (green) subunits form the main part of the structure looking like a crab-claw. The two α subunits (I, purple and II, light blue), ω (dark blue) and σ (red) are located at the periphery. Right: The structure of the bacterial RNA polymerase with the downstream DNA and the DNA-RNA hybrid in the active site (PDB: 205I). The non-template DNA is green, the template DNA is blue and the RNA strand is red.

The jaw-lobe of the claw contains parts of the β subunit (124 kDa) and the shelf has parts of the larger β' subunit (171 kDa). These large subunits are built of many domains each (Figure 10.22). The interface between these two subunits is extensive in the core module. The β subunit embraces the β' with a flap that is flexible. A domain of β' makes a similar interaction with the β subunit. The two α subunits have two domains, and a dimer of their N-terminal domains is located in the core module and further enforces the interaction. The β and β' subunits interact with one α subunit each. The β and β' subunits jointly form the active site. The ω subunit wraps around the C-terminal tail of β' without contact with the active site.

The active site is identified at the bottom of the cleft by a magnesium ion bound to three absolutely conserved aspartate residues.

10.3.2 The σ factor

Most σ factors form a homologous family, the σ 70 family. They are involved in the transcription of thousands of "housekeeping" genes¹ in a bacterial cell. E. coli has six different σ 70 factors, B. subtilis has 18 and S. coelicolor has 63. These different σ factors have specificity for different DNA motifs. On its own, σ cannot bind to the promoter, but bound to the rRNA polymerase it undergoes significant conformational changes and can identify the -35 as well as the -10 elements of the DNA and form the transcription bubble to allow the rRNA polymerase to start transcription. After the initiation phase of transcription, the σ factor dissociates from the RNAP core.

¹Genes that are always expressed in all cells.

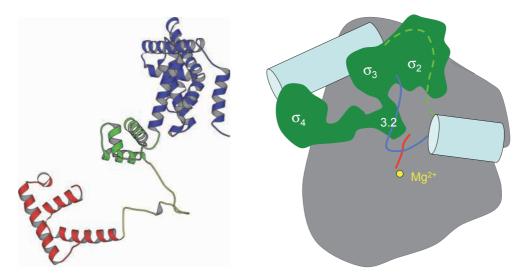


Fig. 10.23 ■ *Left:* The structure of σ 70 from *T. thermophilus* when bound to the core of RNAP. The structure lacks the disordered N-terminal 73 residues. The structure is entirely based on α-helices and flexible loops. The σ_2 domain has eight α-helices (blue); σ_3 has three helices (green); the linker domain 3.2 (yellow) lacks secondary structure; and the C-terminal domain, σ_4 , has four helices (red). *Right:* The structure of *E. coli* RNAP (gray) in complex with σ (green) and the promoter region of DNA (light blue). σ_2 binds to the -10 region and σ_4 has affinity for the consensus sequence of the –35 region. A helix-turn-helix motif in σ_4 , similar to those found in many TFs binds to this region of the DNA (PDB: 1IW7).

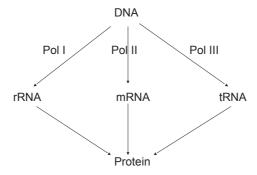


Fig. 10.24 ■ A simplified scheme of the roles of the three different RNA polymerases in eukaryotes.

The σ factor has an elongated U-shaped structure (Figure 10.23). The main part of the protein is composed of four domains built of helices and connected by protease-sensitive flexible linkers. The domains are denoted 1.1, σ 2, σ 3 and σ 4. They contain regions of conserved sequence. Conserved sequences are involved in binding the σ factor to both the rRNA polymerase and the DNA to be transcribed. The flexible linkers allow substantial

conformational changes during the functional cycle. Different σ factors sometimes lack certain elements of the structure. Each domain binds to unique parts of the DNA. $\sigma4$ interacts with the -35 motif, $\sigma 3$ interacts with the extended -10 motif and $\sigma 2$ interacts with the -10 and discriminator motifs.

The σ factor binds to the outer part of the β' half of the claw structure. The $\sigma 4$ domain makes the most extensive contacts with the RNAP core. The U-shaped molecule folds around a part of the β' subunit and the outer surface of the U interacts with N-terminal non-conserved parts of the β' subunit.

For initiation, the transcription bubble is formed. The process of transcription is very similar in bacteria and eukaryotes and is described in Section 10.5.

Eukaryotic Transcription

In eukaryotes, there are three rRNA polymerases: Pol I, II and III (Figure 10.24). Pol I transcribes only one gene containing both of the large ribosomal RNAs. All protein-coding genes are transcribed by Pol II. Pol III transcribes small RNAs such as tRNAs and ribosomal 5S rRNA. Pol I, that transcribes about 60% of cellular rRNA, has 14 subunits. Pol II has 12 and Pol III has 17 subunits. The subunits ($\alpha_2\beta\beta'\omega$) of the core of bacterial rRNA polymerases correspond to subunits of all eukaryotic RNA polymerases (Table 10.2).

rRNA polymerase II (Pol II) has two subunits (Rpb4/7) that form a heterodimer called the polymerase stalk that can dissociate from the core. In Pol I and III, protein complexes A14/43 and C17/25 form the corresponding stalk. Pol I also has the additional complex A49/34.5 and Pol III has C37/53 and C82/34/31, which relates to a GTF, TFIIF in Pol II.

Pol II is the most regulated of the three rRNA polymerases, and there are hundreds of activators and repressors (see Section 10.2). These gene-specific TFs bind to elements in

| TABLE 10.2 Relationship of Conserved Subunits between Different RNA Folymerases | | | | | | | | |
|---|--------|----------|-----------------------|------------------------|---------------|---|---|--|
| Form of RNAP | Genera | lly Cons | erved S | Common | Non-conserved | | | |
| Bacterial ^a | β΄ | β | α_{I} | α^{II} | ω | _ | _ | |
| Archaeal | A'+A'' | В | D | L | K | | 7 | |
| Eukaryotic | | | | | | | | |
| Pol I | A190 | A135 | AC40 | AC19 | Rpb6 | 5 | 4 | |
| Pol II | Rpb1 | Rpb2 | Rpb3 | Rpb11 | Rpb6 | 5 | 2 | |
| Pol III | C160 | C128 | AC40 | AC19 | Rpb6 | 5 | 7 | |

TARLE 10.2 Relationship of Conserved Subunits Retween Different RNA Polymerases

^a Bacterial α subunits (α^I and α^{II}) are identical, but correspond to different proteins in the archaeal and eukaryotic polymerases.

the DNA (regulatory or enhancer elements) that can be up to 50 kb away from the gene. In addition, the GTFs bind to control elements of the promoter region of the DNA close to the gene (Figure 10.1). They recognize the promoter and bind the rRNA polymerase to the promoter but also have additional functions. These GTFs (see Section 10.5) are called TFIIA, B, D, E, F, G/J, H, I and S (TFII identifies a molecule as being a GTF to eukaryotic rRNA polymerase II). The GTFs that are indispensable for initiation are TFIIB, D, E, F and H. One important protein is the TATA-box binding protein (TBP), which is part of TFIID. The binding of TBP to the promoter leads to the formation of a multi-subunit pre-initiation complex (PIC). In some TATA-less promoters, TBP-related factors bind to the promoters.

After transcription, eukaryotic mRNAs also have to be capped and polyadenylated, and the introns must be removed through splicing (see Section 10.7). The processed mRNAs are transported from the nucleus to the cytoplasm. The lifespan of mRNA molecules varies dramatically, which is an important regulatory checkpoint.

10.4.1 Structure of rRNA Polymerase II

The eukaryotic rRNA polymerase II (Pol II) has a molecular mass of around 500 kDa. The structure determination of the yeast enzyme was difficult, taking Kornberg and co-workers about 20 years before their work yielded a well-resolved structure (Figure 10.25). The structures of rRNA polymerases show that five subunits, present in all forms of the enzyme, generate the architecture of the core. Within a radius of about 40 Å from the active site magnesium ion, the enzymes are very similar, but at the periphery their structures differ. Due to this similarity, we will describe the structure of Pol II and comment on deviations in other rRNA polymerases when found important.

The two largest subunits, Rbp1 and Rbp2 of yeast Pol II, correspond to β' and β , respectively in the bacterial RNAP (Table 10.2). They form the two jaws above and below the cleft. Subunits 3 and 11, corresponding to the dimer of α subunits in the bacterial enzyme, anchor the connection of the jaws and nucleate the assembly of the subunits Rpb1 and Rpb2, just as in the bacterial case. The folds of Rpb1 and Rpb2 are unique and interact in a number of ways. At the active site, they are combined to give a single fold. A summary of the structural elements in the polymerase is shown in Table 10.3. The core of the enzyme contains regions of Rpb1 and Rpb2 and the subunits involved in the assembly: Rpb3 and Rpb10-12. The lower jaw contains most of Rpb1 and Rpb5. Rpb2, Rpb9 and a region of Rpb1 form the upper jaw. The clamp, on one side of the cleft, contains mainly the N-terminal regions of Rpb1 and the C-terminal region of Rpb2.

Many of the subunits interact through β -addition motifs where segments of one subunit extend a sheet of a neighbor subunit. Rpb12 that bridges between Rpb2 and Rpb3 participates in two such β -addition motifs. Three Zn²⁺-binding regions stabilize the clamp. In total, eight zinc ions have been identified in Pol II.

TABLE 10.3 Structural Modules and Domains of RNA Polymerases and Their Function

| Name | Subunit in Pol II | Place | Role |
|--------------|----------------------------------|--------------------------|---|
| Upper jaw | Rpb1, Rpb2, Rpb9 | | |
| Lower jaw | Rpb1, Rpb5 | | |
| Jaw-lobe | Rpb1, Rpb9, lobe part of Rpb2 | Upper jaw | |
| Shelf | Rpb1, Rpb5, Rpb6 | Lower jaw | |
| Dock | Rpb1 | | Interacts with the B-ribbon of TFIIB. |
| Wall | Rpb2 | Back wall of active site | Forces the template strand into a 90° different direction for catalysis and formation of the DNA-RNA hybrid. Binds the first cyclin fold of TFIIB. |
| Clamp | Rpb1N, Rpb6, Rpb1C, Rpb2C | | The clamp is a mobile part of the enzyme and interacts with duplex DNA and DNA-RNA hybrid. |
| Rudder | Rpb1 | Loop of the clamp | Separates the DNA-RNA hybrid. |
| Lid | Rpb1 | Loop of the clamp | Separates the DNA-RNA hybrid. |
| Zipper | Rpb1 | Loop of the clamp | The two DNA strands can reunite after the zipper. |
| Fork loop 1 | Rpb2 | | Separates the DNA-RNA hybrid. |
| Fork loop 2 | Rpb2 | | Blocks the continued path of the non- template strand of DNA to initiate the transcription bubble. |
| Bridge helix | Rpb1 | | In the active site. Supports the template base in the +1 position. |
| Trigger loop | Rpb1 | | Flexible loop. Part of the binding site for the substrate nucleotide. Forms a pair of helices as a response to correct substrate recognition, which reduces the width of the entrance funnel. |

The surface of Pol II is almost entirely negatively charged; except for in the active site cleft, that is lined with positively charged residues. A helix of Rpb1 bridges the cleft between the jaws (the bridge helix; Figure 10.25). A loop of the clamp is called the "rudder" and participates in separating the rRNA from the DNA in the hybrid. Three modules of the enzyme change position relative to the core at different states during catalysis. They lie along the sides of the DNA-binding cleft.

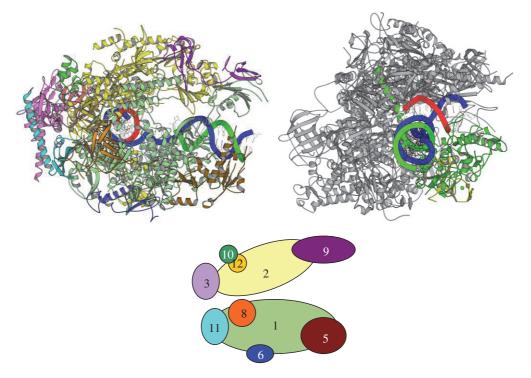


Fig. 10.25 ■ The organization of the core of yeast RNA polymerase II shown in two views. *Top left*: All 10 subunits are individually colored according to the simplified drawing at bottom. *Top right*: A view down the downstream DNA is shown. The bridge helix and the clamp are highlighted in green (PDB: 2E2H). The non-template DNA is green, the template DNA is blue and the RNA strand is red. *Below*: A simplified representation of Pol II in the same orientation as in *Top left*.

10.5 Initiation of Transcription

A central step of initiation of transcription is to accurately bind the DNA to the polymerase and to form the transcription bubble to facilitate access to the template strand (Figure 10.26). Furthermore, the transcription start site (TSS) needs to be identified. The 12-subunit complex of Pol II, which also contains subunits Rpb4/7 has a closed conformation that leaves room only for a single-stranded DNA at the active site. The GTFs (Table 10.4) participate to place the template in the active site. The main knowledge about initiation of transcription comes from studies of promoters with a TATA box. The melting of the dsDNA to form the bubble starts about 12 base pairs downstream of the TATA box. The details of initiation will be described in conjunction with the description of the GTFs.

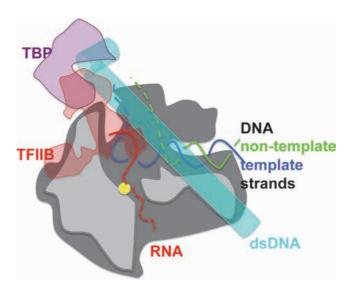


Fig. 10.26 \blacksquare The double-stranded DNA (light blue) to be transcribed by Pol II initially binds to the surface of the enzyme (grey). It gets bent by about 90° by TBP (purple) at the TATA box. This is called the closed promoter complex. TFIIB assists in placing the template strand of DNA into the active site to form the transcription bubble of the open promoter.

TABLE 10.4 The GTFs of RNA Polymerase II

| Name | Number of Subunits | Total Mass kDa | Conserved | Role |
|----------|--------------------|-------------------|--|--|
| TFIIA | 2–3 | 50 | Not in Pols I and III | Assist TBP-binding to TATA-box |
| TFIIB | 1 | 40 | Corresponds to bacterial factor σ | Central in PIC formation. |
| TFIID | 1 | 30 | | TBP binds to TATA box. |
| TFIID | 12–13 | 880 | | TAFs assists TBP in recognition of promoter in particular in TATA-less promoters. |
| TFIIE | 2 | 90 | Small subunit not found in archaea | Interacts with TFIIF and is required for TFIIH binding. |
| TFIIF | 2–3 | 160 | Not found in archaea | Recruits Pol II to the promoter. Regarded as subunits of Pol I and III. The first protein to bind to Pol II during initiation. |
| TFIIH | 11 | 530 | Not found in archaea | Helicase activity to open promoter DNA. Kinase activity. |
| Mediator | 25–30 | 1000 | | Adaptor for transcriptional activation and repression. |

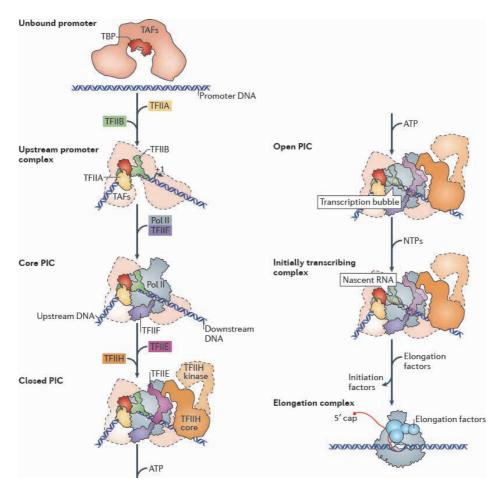


Fig. 10.27 ■ The interaction of GTFs with a DNA promoter and Pol II. (Reproduced with permission from Sainsbury S, *et al.* (2015). Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**: 129–143. Copyright (2015) MacMillan Publishers Limited.)

The key steps of initiation of eukaryotic transcription are presented below (see Figure 10.27):

- (1) A gene-specific activator protein (see Section 10.2) identifies the regulatory motif of a gene. It interacts with histone octamers associated with the gene and with chromatin remodelers (see Section 10.1). The histones are removed or moved so that the gene becomes accessible for transcription.
- (2) TBP identifies the promoter assisted by the gene activator protein. TBP binds to the TATA box and bends the DNA.
- (3) The TAFs of TFIID interact with flanking sequences of the bent DNA at the TATA box.

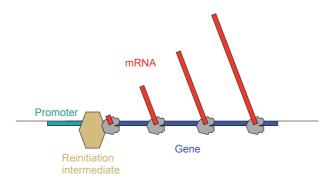


Fig. 10.28 ■ The gene-specific activator protein, the GTFs TFIID, TFIIA, TFIIE, TFIIH and the mediator form an reinitiation intermediate at the promoter, from where several Pol II assisted by TFIIB and TFIIF can initiate several rounds of transcription.

- (4) TFIIB interacts with TFIID and the DNA.
- (5) TFIIF associates with the 12-subunit complex of the Pol II. TFIIB positions the promoter on the polymerase and closes the enzyme-DNA complex. This is the closed promoter complex.
- (6) The complex of Pol II and DNA undergoes a large conformational change that leads to the separation of the DNA strands and the placement of the template DNA strand in the active center of Pol II (Figure 10.26) forming the transcription bubble and the open promoter complex. In this step, TFIIE, TFIIF and TFIIH play major roles by inducing torsional strain in the DNA and thereby promoting strand separation. This step requires ATP and the complex is called the pre-initiation complex (PIC).
- (7) rRNA synthesis begins.

An activated gene is used to initiate multiple rounds of rRNA synthesis. New Pol II molecules can initiate from a reinitiation intermediate (Figure 10.28). This includes gene-specific activator proteins, TFIID, TFIIA, TFIIE, TFIIH and the mediator. On the other hand, TFIIB and TFIIF dissociate and have to participate in each new initiation.

10.5.1 General Transcription Factors

rRNA polymerases need a range of GTFs to produce the mRNAs. These factors are primarily required for initiation and most are multi-subunit complexes (Table 10.4). PIC is formed by Pol II with TBP, TFIIA, TFIIB, TFIIF, TFIIE and TFIIH at the TATA-box.

10.5.1.1 TFIIF

TFIIF is composed of two rRNA polymerase associated proteins called RAP30 and RAP74 in humans. In yeast, they are called Tfg2 and Tfg1, respectively. Both have globular domains

at the termini connected by flexible linkers. The N-terminal dimerization domains of the two proteins bind to the Pol II lobe and the winged helix domains (WH) are located near the DNA downstream of the TATA-box. TFIIF stabilizes the transcription bubble and the interaction between Pol II and TFIIB.

10.5.1.2 TFIID: TBP and TATA box

TFIID is a multiprotein complex including the TATA-box binding protein (TBP) and <u>TBP</u> associated <u>factors</u> (TAFs). TFIID is important for promoter recognition in the DNA and triggers the formation of PIC by providing a scaffold that other parts of the transcription machinery assemble around during initiation. It interacts with transcriptional activators and can read epigenetic marks on the histone tails of nucleosomes. For some TATA-less promoters, TAFs are engaged in promoter recognition. Most genes in yeast depend on TFIID.

The TATA box is an eight base pair part of the promoter usually located about 30 nucleotides before TSS. The consensus sequence is TATA(A/T)A(A/T)(A/G). TBP is a central component for the assembly of the PIC. TBP is ubiquitous, used by all three eukaryotic rRNA polymerases and has a molecular mass around 30 kDa. The DNA-binding C-terminal part of the protein has the shape of a saddle and is built of two domains of 80–90 amino acid residues with very similar structures (Figure 10.29). TBP sits astride the DNA and an eight-stranded antiparallel β -sheet binds to the minor groove of

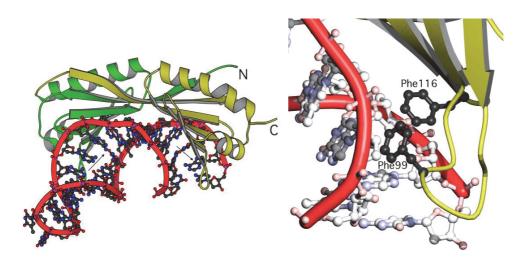


Fig. 10.29 ■ The TATA box binding protein (TBP) from yeast in complex with a DNA hairpin structure. *Left*: The protein has two domains of very similar structure (olive and green). The protein has the shape of a saddle that can bind to double-stranded DNA. It binds to the minor groove of the TATA box and bends the DNA about 90° towards the major groove. *Right*: Two phenylalanine side chains insert between the base pairs. A similar interaction occurs with two phenylalanine residues in the other domain (PDB: 1YTB).

TBP contacts the DNA primarily through non-specific hydrophobic contacts with the bases. The DNA of the TATA box has an inherent tendency of bending and widening the minor groove, which induces TBP to bind. Basic residues of TBP bind to the phosphate groups of the DNA. One important part is the intercalation from the minor groove by two pairs of phenylalanines with the DNA bases at both ends of the TATA box (Figure 10.29). These intercalated phenylalanine side chains lead to a sharp bend of the DNA. The bending of the DNA brings distant parts of the DNA closer to each other so that other TFs can bind upstream or downstream of the TATA box.

10.5.1.3 TFIID: TBP associated proteins (TAFs)

TFIID is essential for cell viability. TFIID is formed by 13 different subunits (TAFs) in addition to TBP. The TAFs contribute to the specificity of TFIID for the promoter. Six of them (TAF4, 5, 6, 9 and 12) are present in two copies each. Nine of the TAFs have domains with the same fold as histones. The histone fold domains of TAFs 6 and 9 form a heterotetrameric complex like histones H3 and H4 (Figure 10.30), and the corresponding domains of TAF4 and TAF12 form pairs. However, an octamer arrangement like in the nucleosome has not been identified. A core complex of TFIID consisting of TAF4, 5, 6, 9 and 12 form a two-fold symmetrical structure (Figure 10.31), but by adding TAFs 8 and 10, the symmetry gets distorted. TAF5 of the core structure has a β -propeller structure with six blades in its C-terminal region.

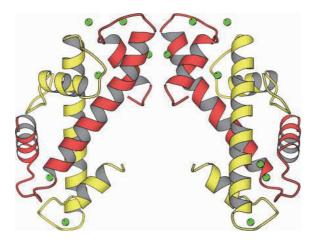


Fig. 10.30 ■ Two dimers of the N-terminal histone fold (HF) domains of TAF6 (yellow) and TAF9 (red) from *D. melanogaster* (PDB: 1TAF). This heterotetrameric complex is similar to the complex between histones H3 and H4 and is part of the core structure of TFIID. The green dots are zinc ions.

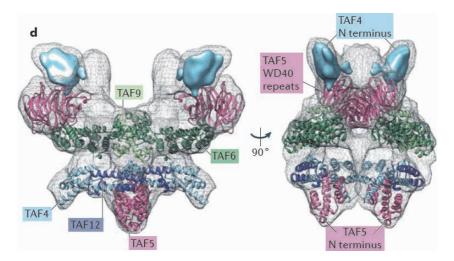


Fig. 10.31 ■ The two-fold symmetric core-structure of TFIID. (Reproduced with permission from Sainsbury S, *et al.* (2015) Structural basis of transcription initiation by RNA polymerase II. *Mol Cell Biol* **16**: 129–143. Copyright (2015) MacMillan Publishers Limited.)

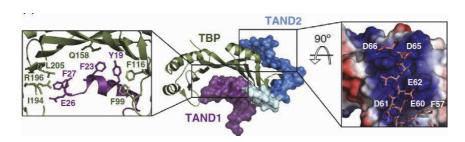


Fig. 10.32 ■ The N-terminal domains TAND1 (purple) and TAND2 (blue) of TAF1 bound to TBP (green) with linker (light blue). TAND1 binds to the concave side of TBP where the TATA-box DNA normally binds. (Reproduced with permission from Kandiah E, *et al.* (2014) More pieces to the puzzle: Recent structural insights into class II transcription initiation. *Curr Opin Struct Biol* **24**: 91–97. Copyright (2014) Elsevier).

TAF1 is the largest of the TAF proteins, with around 1900 amino acid residues in humans. It interacts strongly with TBP through two of its N-terminal domains TAND1 and TAND2. TAND1 mimics the TATA-box DNA and occupies its binding site on the concave side, while TAND2 binds to the convex side of TBP (Figure 10.32). Contrary to expectations, TAND1 functions as an activator and TAND2 as a repressor. Many different proteins are found to interact with the DNA-interacting surface that TAND1 binds to. The dynamics of these interactions may shield critical surfaces of TBP from unproductive complexes to retain access for the proper interaction with DNA and TFs. Several TAFs interact with a range of promoter elements in the DNA. They also have domains that interact with modified lysines or arginines of histones in the nucle-

osomes. Such interactions are found to be important for transcription of TATA-less promoters.

10.5.2 TFIIB

TFIIB is the only GTF that is a single subunit. The polypeptide has several distinct domains or regions, which are important for the formation of PIC. Its role is to identify the transcription start site (TSS) and to bind TBP, TFIID and the bent DNA to the complex between Pol II and TFIIF. Together they form a minimal initiation complex. In archaea TBP and the TFIIB homologue are the only initiation factors needed.

The C-terminal region (TFIIB_C) is composed of two compact domains, both built of five helices each and resembling cyclin A. The sequence identity between the two domains is around 20%. There is no simple two-fold relation between the two domains. TFIIB_C interacts with TBP and the bent DNA on both sides of the TATA box. The interaction sites on the DNA are called the \underline{B} recognition elements (BRE) and may determine the direction or polarity of transcription. The protein only forms limited contacts with the minor or major groove edges of the DNA, making the interaction non-specific.

The role of TFIIB is to bind the complex of TFIID with the TATA box to Pol II in such a way that the DNA can interact with the active site cleft of Pol II. The N-terminal part of TFIIB begins with a zinc binding B-ribbon, followed by the most conserved parts of TFIIB, called the B-reader and B-linker that interact with Pol II (Figure 10.33).

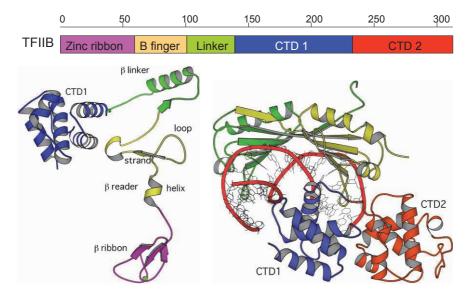


Fig. 10.33 \blacksquare *Top*: The organization of the single polypeptide of TFIIB. The two domains with cyclin fold are in the C-terminal region (TFIIB_C). The B reader in the N-terminal region is the most conserved part of TFIIB. *Bottom left*: The structure of TFIIB (PDB: 4BBR). *Bottom right*: The interaction between TBP and TFIIB_C with a 16 base pair region at a TATA box. TFIIB_C binds at the C-terminal stirrup of the saddle-shaped TBP.

The B-ribbon recruits Pol II by binding at the dock domain of Rpb1 near the path of rRNA exit from Pol II and the first cyclin domain binds at the wall domain of Pol II. The polypeptide of TFIIB following the B-ribbon binds at the exit path for the transcribed rRNA. The B-linker interacts with the DNA where the transcription bubble starts and may participate in the opening and stabilization of the bubble (Figure 10.26). The B-reader helix and loop contacts the template strand of DNA and is stabilized by contacts with the Pol II lid structure (Table 10.3). The B-reader positions the template for the initiation of rRNA synthesis. Pol II undergoes conformational changes upon binding of TFIIB, by which the clamp adopts the closed position. When the rRNA strand has reached 5 nucleotides the B-reader blocks the path and must move. When the rRNA has reached a length of 12–13 nucleotides it clashes with the B-ribbon and TFIIB must be released.

TFIIB and the σ factor superpose in many of their interactions with the rRNA polymerase even though there is no sequence or structural correspondence.

10.5.3 TFIIE and TFIIH

TFII E and TFIIH are needed for opening of the promoter of the DNA. TFIIE assists in the binding of TFIIH to Pol II. TFIIE is a heterodimer and binds to the Pol II clamp domain (Figure 10.34). TFIIH is composed of 10 subunits. These subunits can be divided into two groups, the core consisting of six subunits and the kinase module consisting of three subunits. The tenth subunit links the two units. TFIIH is the only GTF with several enzymatic activities. The kinase activity of TFIIH, by a subunit called CDK7, phosphorylates the C-terminal domain of Rpb1. Two subunits of TFIIH, XPB and XPD (Ssl2 and Rad3 in yeast) are helicases with ATPase activity (Figure 10.34). The XPB activity is needed for

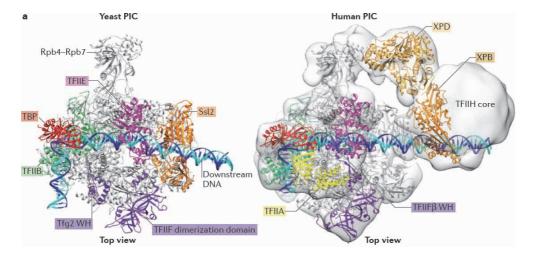


Fig. 10.34 ■ The binding of the closed promoter DNA to Pol II together with GTFs. (Reproduced with permission from Sainsbury S, *et al.* (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**: 129–143. Copyright (2015) MacMillan Publishers Limited.)

10.5.4 The Mediator

The Mediator was discovered as late as in the early 1990s. In yeast, it is a complex of 25 subunits and has a total molecular mass around 1 MDa. It functions as a coactivator both for transcriptional activation and repression. The Mediator functions as an interface between gene-specific DNA-binding transcriptional activators and the general transcription apparatus during initiation and enhances the transcription.

The Mediator is organized as four distinct modules: head, middle, tail and kinase. The first two form the core. The modules all have unique roles. Structural studies of the Mediator have provided crystal structures of the head module and various subunits as well as an EM structure of the Mediator core 15 subunits bound to an initiation complex of Pol II. The Mediator has an elongated shape of about 240 Å and binds to the jaws of the head close to TBP and the middle module that extends beyond Pol II. The Mediator primarily interacts with the C-terminal domain of Rpb1, the largest subunit of Pol II.

10.6 The Active Site of rRNA Polymerases and Transcription

The catalytic mechanisms of bacterial and eukaryotic rRNA polymerases are closely related and can be described together.

10.6.1 Initiation

The promoter of the dsDNA initially binds to the surface of the enzyme and subsequently melts, placing the template strand of the transcription bubble (Figure 10.20) at the active site metal ion, in the cleft between Rpb1 and Rpb2 (β ' and β in bacteria). The downstream duplex DNA is also located in this cleft. The clamp (Table 10.3) forms one side of the cleft and interacts with the ssDNA in the active site and the dsDNA in the downstream region (Figures 10.21 and 10.25). The clamp undergoes a dramatic conformational change in going from an open state without bound DNA to a closed state by a 30° rotation with bound DNA. During transcription, the rRNA forms a hybrid of 9 base pairs with the copied DNA strand in the transcription bubble (Figure 10.20). This hybrid cannot extend linearly from the double-stranded DNA since the wall (Table 10.3) blocks the path. The

RNA-DNA hybrid rather continues in a perpendicular direction (Figures 10.26 and 10.27). Beyond the hybrid region, the DNA and rRNA strands separate and the two DNA strands can reunite assisted by elements of the active site (Table 10.3 and Figure 10.27).

The transcription bubble of the template DNA begins at position +1 in the DNA just at the active site where the DNA-RNA hybrid begins about 30 base pairs from the TATA box. Fork loop 2 of Rpb2 interferes with the direction of the non-template strand of DNA (Figure 10.35). Under some conditions, the base pairs at positions +2 and +3 can also be part of the bubble. The orientation of the DNA base at position +1 is flipped by 90° to form the beginning of the hybrid helix. This orientation is partly due to the bridge helix. Fork loop 1, the lid and the rudder separate the rRNA from the template strand.

10.6.2 Elongation

The rRNA polymerases go through three states during the elongation phase of transcription: the pre- and post-translocational and the backtracked states. In the first one, the nucleotide just added to the rRNA transcript still occupies the addition site. In the post-translocational state this site is empty, ready to bind a new NTP matching the new DNA nucleotide in position +1. In the backtracked state, the DNA-RNA hybrid moves backwards one or several bases with the rRNA moving into the funnel (Figure 10.35).

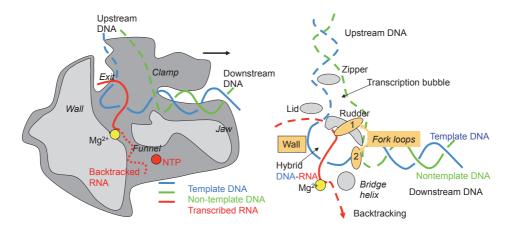


Fig. 10.35 ■ *Left*: The functional components of Pol II. The jaws are seen on the lower right and the clamp above the duplex DNA. The hybrid DNA-RNA helix is forced upwards by the wall. The magnesium ion (yellow) corresponds to metal A. Nucleotides (NTP) enter into the active site through the funnel below. *Right*: A number of features of the enzyme guide the nucleic acids. Fork loop 2 participates in splitting the duplex DNA into the transcription bubble. The template strand enters the active site, but the non-template strand is led a different route. Fork loop 1, the lid and rudder split the hybrid helix to let the duplex DNA reform after passing the zipper.

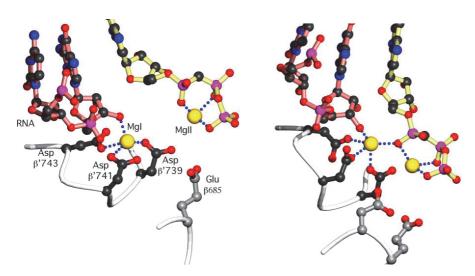


Fig. 10.36 ■ *Left*: The nucleotide in the preinsertion site observed in crystals of *T. thermophilus* RNAP inhibited by streptolydigin. The incoming nucleotide (AMPcPP with yellow bonds) base pairs with the template DNA in position +1, but the α-phosphate is too far from the growing RNA chain to be incorporated (PDB: 2PPB). *Right*: The nucleotide (AMPcPP) bound without the inhibitor. The two magnesium ions are now close. A bridging water molecule (not shown) fitted to the missing coordination position for both metals could participate in catalysis and hydrolyze the pyrophosphate off (PDB: 2O5J).

In the bacterial system, in addition to the insertion site, a preinsertion site has been identified when an inhibitor (streptolydigin) was included in the system (Figure 10.36). The incoming nucleotide (AMPcPP) is base-paired to the DNA nucleotide in position +1. However, in the preinsertion site the α -phosphate of the substrate nucleotide is too far away from the O3' hydroxyl for catalysis. Two magnesium ions are seen, called MgI (or metal A) and MgII (or metal B). MgI is bound to the bridging phosphate of the last two nucleotides of the transcribed rRNA and to conserved aspartate residues of a loop of Rpb1. MgII is bound to the three phosphates of the nucleotide in the preinsertion site and aspartate residues from both Rpb1 and Rpb2.

When the inhibitor is left out, a different structure is seen (Figure 10.36). Here, the AMPcPP molecule is properly placed in the insertion or A-site. The pyrophosphate can then be hydrolyzed off from the NTP and the nucleotide incorporated into the growing mRNA.

The enzyme undergoes interesting conformational changes comparing the nucleotide binding to the preinsertion and insertion sites where the bridge helix is an essential component. In the post-translocation state, with the empty A-site, the active site is in an open state with a mobile trigger loop. Upon binding of a cognate NTP the trigger loop folds into a helical hairpin, closing the active site (Figure 10.37). Binding of a non-cognate NTP does not induce this conformational change, which simplifies the NTP release. Together with the

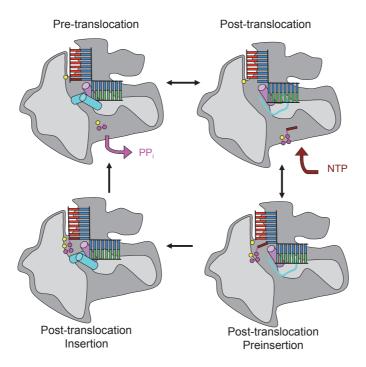


Fig. 10.37 ■ The elongation cycle of transcription. The template strand of DNA (blue) and the non-template DNA (green) are bound to the RNA polymerase. The RNA (red) makes a hybrid arrangement with the template DNA. The bridge helix (purple), the trigger loop (light blue) and Metal A (yellow) are shown. *Top right*: An NTP molecule (brown base with three purple phosphates) and Metal B (yellow) are recruited through the funnel to base pair with the template DNA in the +1 position, at the active site. *Bottom right*: The NTP is bound at the preinsertion position. *Bottom left*: The NTP is incorporated into the A-site. The trigger loop changes its conformation to extend both neighboring helices to become the trigger helices. *Top left*: The catalytic step of the enzyme. A cognate ribonucleotide is incorporated into the growing rRNA chain. The pyrophosphate group and MgII dissociate. In the subsequent step, the next DNA nucleotide is placed in the active site to be transcribed.

bridge helix, the trigger loop forms a three-helical bundle. Formation of the trigger helices reduces the dimensions of the entrance of the funnel into the active site. This closed conformation is the active conformation of the enzyme. The inhibitor streptolydigin prevents the conformational transition of the trigger loop and thus to the active conformation of RNAP.

10.6.3 The rRNA-DNA Hybrid Helix and Backtracking

When the transcripts reach a length of 10 residues, the rRNA is separated from the template strand of the DNA and the two DNA strands can reunite. Three peptide loops extending from the clamp, called the "rudder", "lid" and "zipper", play roles in the

dissociation of the hybrid (Figure 10.35). The lid is not particularly conserved and does not interact strongly with the hybrid. It plays a steric role in RNA-DNA strand separation. The rRNA structure beyond the hybrid is single-stranded but remains stacked. The rRNA chain is threaded towards the surface of the enzyme.

rRNA polymerases oscillate between forward (polymerization) and backward (backtracking) movements during transcription. Mutations of the trigger loop enhance either forward translocation or backtracking. The backtracking is used during initiation, for proofreading and when there is DNA damage. Transcription normally begins with repeated attempts to synthesize short RNAs ("abortive attempts") until a limit of about 10 nucleotides is passed. Below this length, the rRNA is released from the hybrid. When the transcript reaches the size of 20 nucleotides, it becomes fully stable.

The specificity for ribo- rather than deoxy-ribonucleotides is probably due to recognition of both the ribose sugar and the DNA-RNA hybrid helix. The protein is highly complementary to the non-standard conformation of the hybrid helix. A deoxyribonucleotide or an incorrect base in the region -1 – -5 will be destabilizing and lead to backtracking. In such a case, the 3'-end of the rRNA is released from the hybrid and the RNA-DNA hybrid is temporarily reformed at the 5'-end. The funnel through which NTPs enter into the active site is also the pore through which rRNA is extruded in the case of backtracking. The misincorporated nucleotide can then be removed by cleavage at the active site by GTF TFIIS.

10.7 Splicing

The eukaryotic genes, and therefore the transcribed precursor pre-mRNAs, contain regions which are not translated. This is illustrated in Figure 10.1, where the exons are the regions that will be kept and the introns are the regions that will be removed through a process called splicing, usually performed by a large molecular complex called the spliceosome composed of a number of rRNA and protein molecules. Other types of splicing are the introns involved in self-splicing described in Chapter 5. In splicing, the pre-mRNA is cut at specific sites at the 5'-end and the 3'-end of the intron.

10.7.1 The Spliceosome

In budding yeast, splicing involves five small nuclear RNAs (snRNAs) and around 100 different proteins, whereas in mammals nine snRNAs and over 300 different proteins participate. The snRNAs in the smaller spliceosomes are called U1, U2, U4, U5 and U6, each associated with a number of specific proteins and therefore called snRNPs. The main mass of the spliceosome is due to the many proteins.

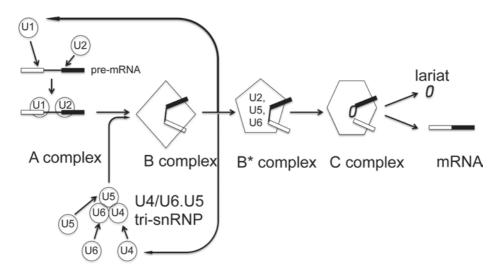


Fig. 10.38 ■ A simplified representation of the process of spliceosome assembly and activation in yeast.

The assembly and disassembly of the spliceosome is complex (Figure 10.38). U1 and U2 snRNPs recognize the pre-mRNA 5' splice site and form the A complex. The preformed tri-snRNP complex (composed of the snRNPs U4, U5 and U6) binds to the A complex and forms the B complex. Then the catalytically active B* complex forms through conformational rearrangements whereby U1 and U4 dissociate. Subsequently, an rRNA lariat is formed of the intron in the C complex. The lariat is subsequently released and the exon parts of the pre-mRNA are joined to form the mRNA.

Due to structural investigations, the complex structure and mechanism of the yeast spliceosome have started to emerge. Structures of the tri-snRNP (Figure 10.39, *left*) and a complex containing the lariat are available (Figure 10.39, *right*). The dimensions of the spliceosome are in excess of 300 Å. The central body has a triangular shape from which a head and two arms extend (Figure 10.39, *right*). The variety of complex and extended protein structures is remarkable. The structures and the functional interplay make it evident that the spliceosome is a very dynamic molecular assembly.

The active site is located at the heart of the U2.U5.U6 spliceosome close to the three rRNA molecules. The large protein Prp8 forms the main scaffold for pre-mRNA splicing. It contains domains related to a reverse transcriptase (RT) with palm, fingers and thumb domains. At least two magnesium ions, M1 and M2, are associated with the catalytic reaction. In the first step of splicing, the 2′-OH group of a conserved adenine, in the branch point sequence (BPS) of the intron, is activated by M2 and makes a nucleophilic attack on the phosphorous atom of the guanine nucleotide at the 5′-end of the intron, resulting in the release of the 5′-exon and formation of an intron lariat-3′-exon. M1 stabilizes the leaving group of the 3′-end nucleotide in the 5′-exon. The spliceosome is clearly a ribozyme.

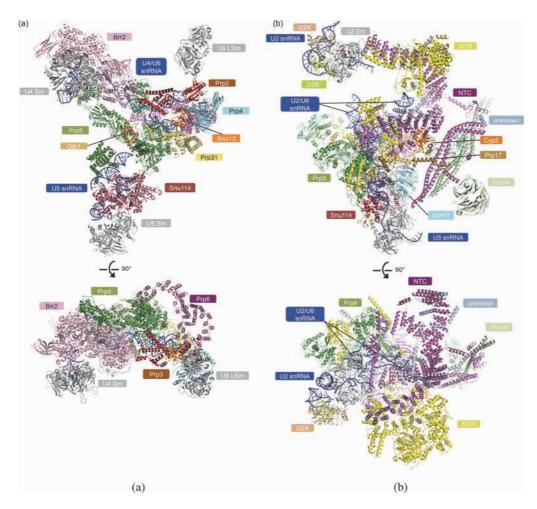


Fig. 10.39 ■ Two structures of different stages of the spliceosome functional cycle. Left: The U4/U6.U5 tri-snRNP. Right: The U2.U5.U6 spliceosome with an intron lariat bound. (Reproduced with permission from Nguyen HD, et al. (2016) CryoEM structures of two splieosomal complexes: Starter and dessert at the spliceosomal feast. Curr Opin Struct Biol 36: 48-57. Printed by Elsevier. Copyright (2016) the authors.)

For Further Reading (Section 10.1)

Reviews

Hughes AL and Rando OJ. (2014) Mechanism underlying nucleosome positioning in vivo. Ann Rev Biophys 43: 41-63.

Mueller-Planitz F, Klinker H and Becker PB. (2014) Nucleosome sliding mechanism: New twists in a looped history. Nat Struct Mol Biol 20: 1026-1032.

Patel DJ and Wang Z. (2013) Readout of epigenetic modifications. *Ann Rev Biochem* **82**: 81–118. Zhou B-R, Jiang J, Feng H, Girlando R, *et al.* (2015) Structural mechanisms of nucleosome recognition by linker histones. *Mol Cell* **59**: 628–638.

For Further Reading (Section 10.2)

Original Articles

- Canadillas JMP, Tidow H, Freund SMV, et al. (2006) Solution structure of p53 core domain: Stuctural basis for its instability. *Proc Natl Acad Sci USA* **103**: 2109–2114.
- Cho Y, Gorina S, Jeffrey PD, Pavletich NP (1994) Crystal structure of a p53 tumor suppressor-DNA complex: Understanding tumorigenic mutations. *Science* **265**: 346–355.

Reviews

Slattery M, Zhou T, Yang L, *et al.* (2014) Absence of a simple code: How transcription factors read the genome. *Trends Biochem Sci* **39**: 381–399.

For Further Reading (Section 10.3-10.6)

Original

- Basu RS, Warner BA, Molodtsov V, *et al.* (2014) Structural basis of transcription initiation by bacterial RNA polymerase holoenzyme. *J Biol Chem* **289**: 24549–24559.
- Bieniossek C, Papai G, Schaffizel C, et al. (2013) The architecture of human general transcription factor TFIID core complex. *Nature* **493**: 699–702.
- Cramer P, Bushnell DA, Fu J, et al. (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* **288**: 640–649.
- Engel C, Sainsbury A, Cheung AC, et al. (2013) RNA polymerase I structure and transcription regulation. *Nature* **502**: 650–655.
- Nikolov DB, Chen H, Halay, ED, et al. (1995) Crystal structure of a TFIIB-TBP-TATA-element ternary structure. *Nature* **377**: 119–128.
- Plaschka C, Larivière L, Wenzeck L, *et al.* (2015) Architecture of the RNA polymerase II Mediator core initiation complex. *Nature* **518**: 376–380.

Reviews

- Feklístov A, Sharon BD, Darst SA, Gross CA. (2014) Bacterial sigma factors: A historical, structural and genomic perspective. Ann Rev Microbiol 68: 357–376.
- Kandiah E, Trowitzsch S, Gupta K, et al. (2014) More pieces of the puzzle: Recent structural insights into class II transcription initiation. Curr Opin Struct Biol 24: 91–97.
- Liu X, Bushnell DA, Kornberg RD. (2013) RNA polymerase II transcription: Structure and mechanism. Biochim Biophys Acta 1829: 2-8.
- Sainsbury S, Bernecky C, Cramer P. (2015) Structural basis of transcription initiation by RNA polymerase II. Nat Rev Mol Cell Biol 16: 129-143.

For Further Reading (Section 10.7)

Original Articles

- Galej WP, Wilkinson ME, Fica SM, et al. (2016) Cryo-EM structure of the spliceosome immediately after branching. Nature http://dx.doi.org/10.1038/nature19316 (2016).
- Hang J, Wan R, Yan C, Shi Y. (2015) Structural basis of pre-mRNA splicing. Science 349: 1191–1198. Nguyen HD, Galej WP, Bai X, et al. (2016) Cryo-EM structure of the yeast U4/U6.U5 tri sn-RNP at 3.7 Å resolution. *Nature* **530**: 298–302.
- Wan R, Yan C, Bai R, et al. (2016) The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. Science 351: 4 66–475.
- Yan C, Hang J, Wan R, et al. (2015). Structure of a yeast spliceosome at 3.6 Å resolution. Science 349: 1182-1191.

Review

Nguyen HD, et al. (2016) Cryo-EM structures of two spliceosomal complexes: Starter and dessert at the spliceosomal feast. Curr Opin Struct Biol 36: 48-57.

Protein Synthesis — Translation

11.1 Evolution of the Translation System

The translation of genetic information into functional protein molecules is a central process for life; and the genetic code, tRNA molecules and the machinery for protein synthesis are all highly conserved. The ribosome on which translation occurs is composed of protein and rRNA molecules. Carl Woese showed that sequenced fragments of ribosomal rRNA molecules from a large variety of species were related. Thus, the ribosomal RNAs could be used to analyze the relationship between species. By 1977, he could show that it was not correct to divide the present living organisms into prokaryotes and eukaryotes. A unique new kingdom had to be introduced: archaea. Living organisms, according to Woese, had to be organized into bacteria, archaea and eukaryotes.

In comparisons of completely sequenced genomes, the molecules of the translation apparatus stand out as dominating the group of universally conserved molecules. The genetic code, tRNAs, ribosomal rRNA, ribosomal proteins and translation factors must have coevolved at a very early phase of biological evolution and have subsequently gone through only limited further changes.

An important aspect of protein synthesis is that nucleic acid molecules have central roles, in contrast to most other processes in cells where proteins dominate. Central components are the mRNA, the tRNA and the ribosomal rRNA molecules. An mRNA molecule contains a copy of the gene sequence and binds to the ribosome. The tRNA molecules, the adapters suggested by Francis Crick, decode the gene sequence and link the amino acid into the growing peptide on the ribosome.

| 2nd base in codon | | | | | | | | |
|----------------------------|---|----------------------------------|-------------------------|----------------------------------|----------------------------------|------------------|-------------|--|
| | | U | С | Α | G | | | |
| 1st base in codon | U | Phe F Phe F Leu L Leu L | Ser S Ser S | Tyr Y Tyr Y STOP STOP | Cys C Cys C STOP Trp W | U C A G | | |
| | С | Leu L Leu L Leu L Leu L | Pro P Pro P | His H His H Gln Q Gln O | Arg R Arg R Arg R Arg R | U C A G | 3rd base | |
| | A | Ile I Ile I Ile I Met M | Thr T Thr T Thr T | Asn N Asn N Lys K Lys K | Ser S Ser S Arg R | U C A G | in codon | |
| | G | Val V Val V Val V Val V | Ala A Ala A | Asp D Asp D Glu E Glu E | Gly G Gly G | U C A G | | |

Fig. 11.1 ■ The universal genetic code. The trinucleotide codons are translated to the 20 amino acids, given here with their three and one letter codes.

11.1.1 The Genetic Code and the tRNAs

The base triplets of the genetic code (Figure 11.1) that build up a messenger rRNA (mRNA) are called codons and correspond to the 20 different amino acids. In addition, there are normally three stop codons (UAA, UAG and UGA). The synthesis of a protein normally starts at an AUG codon, which is also the codon for methionine. Special systems have developed to differentiate between ordinary methionine codons and the start signal, the initiation codon. In some cases (methionine, tryptophan), there is only one triplet that codes for a certain amino acid, but sometimes there are as many as six (serine, leucine, arginine). The code is degenerate. There is no one-to-one relationship between codons and tRNA molecules. Mammalian mitochondria have a very limited set of tRNA molecules, but many species have around 40. This relates to the variable codon usage as well as to the capacity of some tRNAs to read several codons (tRNA wobble base-paring).

The tRNA molecules are normally around 75 nucleotides long (Section 5.3.10.1). The secondary structure looks like a cloverleaf with a stem and three leaves (Figure 5.47). The stem has the unique sequence CCA at the 3'-end. It is the ribose of the 3'-terminal A that gets aminoacylated by specific enzymes, tRNA synthetases. This stem is therefore called the aminoacyl or acceptor stem. The three leaves or arms of the tRNA are called the D stem and loop, the anticodon stem and loop, and the T stem and loop, respectively. In addition, there is a variable loop (V-loop) that can be upto 21 nucleotides in length. Serine and leucine tRNAs generally have long V-loops as does tyrosine tRNA in bacteria and chloroplasts. The anticodon at the end of the middle loop can match a codon in the mRNA. The three-dimensional structure of the tRNA molecule has the shape of an "L" (Figure 5.47). The acceptor stem and the T stem form the long leg of the "L" while the anticodon and D stems form the shorter leg. The T and D loops interact to form the elbow.

Thus, tRNA has a surprising arrangement, with its two functional parts at opposite ends of the molecule, about 75 Å apart.

tRNA Synthetases

The enzymes that charge the tRNAs with amino acids (aa), the aminoacyl-tRNA synthetases (aaRS), are specific for each amino acid. Since there are 20 amino acids, there are normally 20 tRNA synthetases in an organism. The tRNA molecules are charged in a two-step process (Figure 11.2):

- (i) $aa + ATP \longrightarrow aa-AMP + PP$,
- (ii) $aa-AMP + tRNA \longrightarrow aa-tRNA + AMP$.

The correct amino acid has to be bound by the enzyme and activated by an ATP molecule to form the reactive intermediate aa-AMP [step (i)]. In step (ii), the amino acid is transferred to the correct tRNA, that is also bound to the enzyme. The 2'- or 3'-OH group of the terminal adenosine of the conserved CCA motif directly attacks the high-energy ester bond in aa-AMP, resulting in attachment of the amino acid to the ribose. The fidelity of translation depends primarily on the synthetases. Errors in the charging of the tRNAs will not be detected in subsequent steps. Thus, the synthetases must recognize their specific amino acid and the tRNA with high accuracy. Since some amino acids are very similar in size and structure, special editing mechanisms have evolved (Figure 11.2).

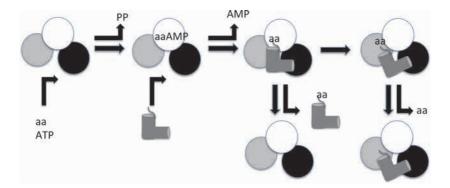


Fig. 11.2 ■ Aminoacylation and editing of a tRNA by a aaRS. The black domain interacts with the anticodon. The white domain is the aminoacylation domain while the gray domain is the editing domain. In the first step, the amino acid (aa) is activated by ATP. In the second step, the tRNA is charged with an amino acid and the charged tRNA is either released or edited.

The molecular weight and the oligomeric state of the aaRS vary considerably (Table 11.1). Two classes of tRNA synthetases have been identified on the basis of the three-dimensional structures and sequence similarities. There are ten aaRS in each class. While the enzymes of class I are normally monomeric, those of class II are always dimers or tetramers.

The enzymes of the two classes have entirely different structures. The aaRS are modular and built of several different domains. While the ATP-binding or catalytic domain of class I aaRS is a Rossmann fold with parallel β -strands (Figures 11.3 and 11.4), the corresponding domain in the class II aaRS is built from antiparallel β -strands (Figure 11.5). Different consensus sequences of the catalytic domain are characteristic for the two classes (Table 11.2). The aaRS of the two classes recognize the tRNAs from opposite sides and charge the tRNA on the 2' OH (class I) or the 3' OH (class II) of the terminal ribose of the tRNA.

Not only do the ATP-binding domains display characteristic features, but also the other domains that build up these enzymes. Thus, the two classes of aaRS can be divided into subclasses a, b and c based on sequence homology and domain architecture

| Table 11.1 | Oligomeric States of | f Aminoacyl-tRNA Sy | nthetases |
|-------------------|----------------------|---------------------|-----------|
| | | | |

| Class I | | | | | | | | | | | |
|------------------|------------|------------|-------------------|------------|------------|------------|------------|------------|------------|-------------------|------------|
| RS | L | I | V | C | M | R | E | Q | K | Y | W |
| Oligomeric state | α | α | α | α | α_2 | α | α | α | α | α_2 | α_2 |
| Class II | | | | | | | | | | | |
| RS | S | T | G | A | P | Н | D | N | K | F | |
| Oligomeric state | α_2 | α_2 | $(\alpha\beta)_2$ | α_2 | α_2 | α_2 | α_2 | α_2 | α_2 | $(\alpha\beta)_2$ | |

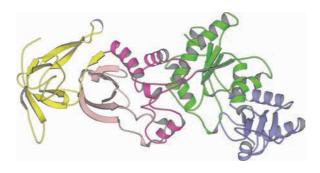


Fig. 11.3 \blacksquare The structure of the Gln synthetase (GlnRS) belonging to class I subclass b. The catalytical domain (green) has the characteristic Rossmann fold with the active site at the C-terminal end of the parallel β-sheet. Domain II, the editing domain (blue), is inserted into the catalytic domain. Domain III (purple) is a helical domain, while domains IV (yellow) and V (pink) are antiparallel β-barrels (PDB: 1GTR).

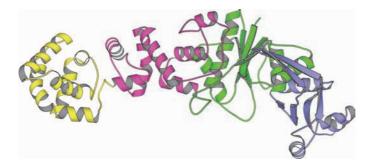


Fig. 11.4 ■ The structure of the Glu synthetase (GluRS). Even though this enzyme belongs to the same class and subclass as GlnRS it has a related but distinctly different structure. For example, the C-terminal two domains (purple and yellow) are both helical (PDB: 1GLN).

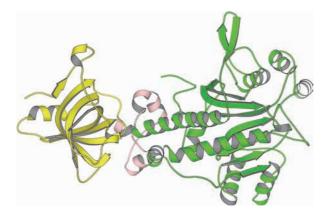


Fig. 11.5 \blacksquare The structure of the Asp synthetase (yeast). AspRS belongs to class II and the catalytic domain (green) is characterized by an antiparallel β-sheet (PDB: 1EOV).

TABLE 11.2 Characteristics of the aaRS Enzymes

| | | Class I | Class II |
|--------------------|---|------------------------|----------------------|
| Sequence motif | | HIGH | FRXE/D |
| | | KMSKS | R/HXXXF |
| | | GXGXGXER | |
| Subclass | a | L, I, V, C, M, R | S, T, G, A, P, H |
| | b | E, Q, K | D, N, K |
| | c | Y, W | F |
| Aminoacylation | | 2'OH | 3'OH |
| Fold of ATP domain | | Rossmann (// β) | Antiparallel β |
| Amino acid binding | | Surface | Deep pocket |
| tRNA acceptor end | | Bent | Straight |
| | | | |

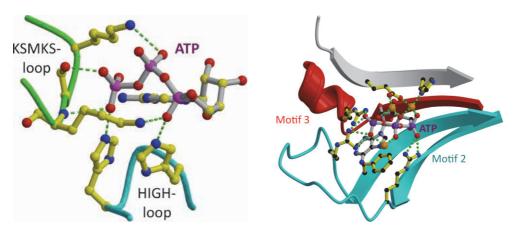


Fig. 11.6 ■ The binding of ATP to one tRNA synthetase from each class. *Left*: TyrRS (class Ic) with conserved residues of the KMSKS loop (green) and HIGH loop (blue) interacting in specific ways with the ATP. Right: ATP binding to ProRS (class IIa). Residues of motifs 2 (blue) and 3 (red) make specific interactions with the ATP. (figure obtained from Stephen Cusack.)

(Figures 11.3–11.5), where the two types of ATP-binding domains define the class and similar domain arrangements define the subclass. However, as shown in Figures 11.3 and 11.4, even for enzymes within one subclass there are significant differences. GlnRS and GluRS have related inserts into the catalytic domain. The structures that follow immediately after the catalytic domain are also related. However, the two C-terminal domains in both enzymes are distinctly different.

The binding of ATP and amino acids to the enzyme needs to be specific. The ATP binds to the characteristic sequence motifs of the two classes (Figure 11.6). In contrast, the binding of the amino acid uses a whole range of different types of interactions (Figure 11.7).

11.2.2 Binding of tRNAs

The identification of the correct (cognate) tRNA among the different non-cognate ones is due to a number of features of the individual tRNAs called the "recognition elements" (Figure 11.8). The tRNA synthetases of the two different classes bind to opposite sides of the tRNA. Most of the recognition elements are localized on the side of the tRNA that faces the aaRS. The recognition of the different tRNAs by the aaRS in most cases involves the acceptor stem and the anticodon (Figure 11.9). The recognition of the middle base (position 35) of the anticodon is very common.

Since the amino acid is attached at the end of the tRNA opposite the anticodon, the synthetases need to be elongated to interact with both ends of the tRNA (Figure 11.9). In the case of the Asp synthetase, the bases of the anticodon interact with the anticodon-binding

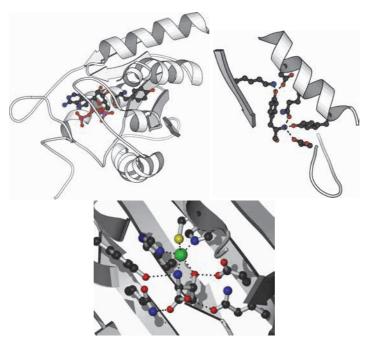


Fig. 11.7 ■ Amino acid binding in synthetases. *Top left*: The catalytic domain of TyrRS (class Ic) with a bound ATP molecule and a tyrosine ready to react (PDB: 1H3E). *Top right*: The tyrosine is specifically recognized by TyrRS through hydrogen bonds to the OH-group. *Bottom*: The binding of threonine to ThrRS (class IIa) involves a Zn ion, making contacts with the hydroxyl group and the amino group (PDB: 1EVK).

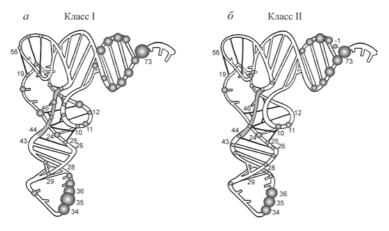


Fig. 11.8 ■ The distribution of recognition elements for tRNA synthetases on tRNAs of class I (*left*) and class II (*right*). The variable arm can be upto 21 nucleotides and is recognized in the case of Ser and Tyr tRNAs. (Reprinted with permission from Vasileva IA, Moor NA. (2007) Interaction of aminoacyl-tRNA synthetases with tRNA: General principles and distinguishing characteristics of the high-molecular-weight substrate recognition. *Biochemistry* (*Moscow*) 72: 306–324. Copyright Springer Verlag.)

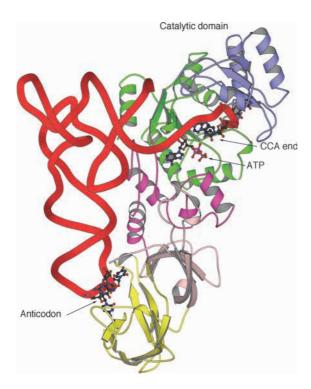


Fig. 11.9 ■ The Gln synthetase (class I) in complex with the corresponding tRNA. Both the acceptor end and the anticodon are in contact with the enzyme (PDB: 1GTR). The proximity of the ATP with the bent CCA end is also evident. The editing domain is blue.

domain and specificity is achieved through several hydrogen bonds. Further interactions through stacking and contacts with the rRNA backbone stabilize the binding (Figure 11.10).

In the case of the leucine and serine (subclass Ia) and alanine and glycine synthetases (subclass IIa), the anticodon is not part of the interaction. For both serine and leucine there are six different codons. For leucine, the second base is always U but for serine all positions of the anticodon nucleotides can be different. Therefore, it would be hard for a single enzyme to discriminate through interactions with the anticodons. Both leucyl and seryl tRNAs instead have a very long variable arm. The complex of the Ser synthetase with tRNA shows that this class II enzyme has an extended helical hairpin, which is inserted between the TYC and variable arms of the tRNA (Figure 11.11). This loop is disordered in the absence of tRNA. The anticodon stem is pointing away from the enzyme, but the acceptor stem is bound close to the active site of the other subunit in the dimer. For the leucyl synthetase the long variable arms of the different tRNAs for leucine are recognized by the archaeal enzyme, but is not directly used by the bacterial enzyme.

For the reaction of aa-AMP with tRNA, the reactants are placed close together to be able to perform the transfer reaction.

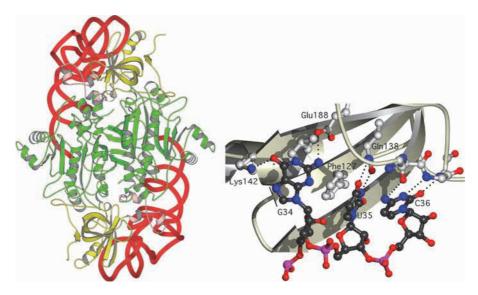


Fig. 11.10 ■ Left: The Asp synthetase dimer (class II) with two bound tRNA molecules. Right: The interaction of the anticodon GUC of Asp-tRNA with conserved residues of the Asp synthetase. Hydrogen bonds (Lys142, Glu188, Gln138, main chain) as well as stacking (Phe127) interactions are seen. The hydrogen bonds are shown by dotted lines. All class IIb synthetases recognize anticodons with a central uracil, recognized by Phe127 and Gln138 (PDB: 1ASY).

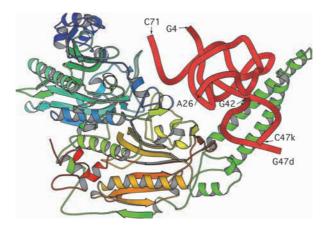


Fig. 11.11 ■ The dimeric SerRS has a long coiled-coil structure (green), which interacts to identify the long variable arm of the tRNA (PDB: 1SER). The anticodon stem and loop (between nucleotides 26 and 42) and the variable arm loop (between 47d and 47k) are not visible in the structure.

11.2.3 Selection of Amino Acids — Editing

Not only the tRNAs may be difficult to differentiate, but also some amino acids are very similar in shape and nature. Thus, valine would fit nicely into a pocket designed for isoleucine, threonine fits into a pocket defined for valine, and serine fits into pockets for threonine. The affinities for the non-cognate amino acids may be lower but this does not give sufficient discrimination against wrong aminoacylation. These amino acid similarities are found within subclasses Ia and IIa. Misacylated tRNAs or misactivated amino acids need to be eliminated. This is done both through pre- and post-transfer editing mechanisms. The post-transfer mechanism found in certain aaRS is best understood. The selection is done through a "double sieve system" (Figure 11.12). A separate editing domain is used to hydrolyze the amino acid from the incorrectly acylated tRNA (Figure 11.13).

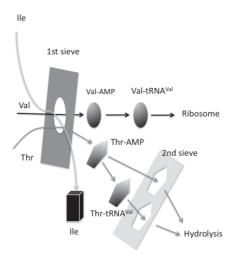


Fig. 11.12 ■ An example of a double sieve mechanism for the selection of the correct amino acid in Val synthetase: the double sieve system to select valine by ValRS.

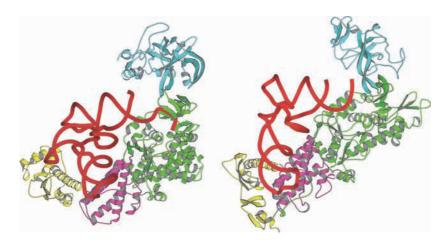


Fig. 11.13 ■ Two class I enzymes with editing domains (cyan). *Left*: LeuRS is an enzyme that does not bind directly to the anticodon. The recognition partly depends on the variable arm, which is longer in tRNA^{Leu} than in most tRNAs (PDB: 1WZ2). *Right*: The similar IleRS illustrates the role of the editing domain. The acceptor arm can bind alternatively in the catalytic and the editing domains. Note that here the anticodon is used for recognition (PDB: 1QU2). The editing domain is similar in these two enzymes and in ValRS, but has a different fold in the editing domains of class II enzymes.

LeuRS, IleRS and ValRS (class Ia) all have a homologous and conserved editing domain called CP1. The editing domains of ThrRS and ProRS, however, are not conserved. In the case of ValRS, the editing domain has a binding site that allows threonine to bind but be hydrolyzed by the enzymatic activity of the editing domain (Figure 11.13). This editing domain-binding site cannot bind and remove a correctly incorporated valine. In the hydrolysis activity of the editing site, the free ribose hydroxyl (3' OH for class I and 2' OH for class II) of the terminal adenine (A76) of the tRNA has an important role.

11.3 The Ribosome

11.3.1 Ribosome Composition and Function

The ribosome is a large complex of protein and rRNA where a messenger rRNA is translated into a sequence of amino acids. The ribosome is composed of a large and a small subunit, which can dissociate and reunite. The small subunit is involved in binding of the mRNA and decoding its message. Peptidyl transfer occurs on the large subunit. In contrast to DNA and rRNA polymerases, which are protein enzymes, most of the essential functions associated with ribosomes depend on rRNA molecules. In bacteria, the small subunit (30S) has one rRNA molecule, the 16S ribosomal rRNA (rRNA). The Svedberg unit, S, is a measure of the sedimentation velocity in the ultracentrifuge. The small subunit has normally 21 proteins (old names S1 to S21). The large bacterial subunit (50S) has two rRNA molecules, 5S and 23S, rRNA, and about 33 proteins (old names L1 to L36; three numbers do not correspond to unique proteins). In addition, the mRNA and the tRNA molecules are essential for protein synthesis on the ribosome.

Archaeal ribosomes and rRNA molecules are of approximately the same size as bacterial ribosomes but the small and large subunits have about 28 and 40 proteins, respectively (Table 11.3). Eukaryotic ribosomes are larger. The small subunit (40S) contains the 18S rRNA and about 33 proteins. The large eukaryotic subunit (60S) contains the 5S, 5.8S, and 28S rRNAs, and about 47 proteins. In mammalian mitochondria, the rRNA molecules are significantly smaller, 12S and 16S, while the number of proteins is larger. About half of the proteins are related between mitochondrial, eubacterial, archaeal and eukaryotic ribosomes. In order to make the naming of ribosomal proteins consistent regardless of species a new convention has been introduced. Half of the ribosomal proteins, which are universal, keep their bacterial name preceded by a u (like uS2 or uL1). Proteins only found in bacteria keep their name but it is preceded by a b (like bS1 or bL12). Proteins only found in eukaryotes keep their old name preceded by an e (like eS1 or eL6). The ribosomal proteins from archaebacteria seem to be fully covered by the universal or eukaryotic names.

| Source | | Size | RNA | Proteins |
|------------------------|---------------|------|---------------|----------|
| Bacteria | | 70S | | |
| | Small subunit | 30S | 16S | 23 |
| | Large subunit | 50S | 23S, 5S | 33 |
| Archaea | | 70S | | |
| | Small subunit | 30S | 16S | 28 |
| | Large subunit | 50S | 23S, 5S | 40 |
| Eukaryotes | | 80S | | |
| | Small subunit | 40S | 18S | 33 |
| | Large subunit | 60S | 28S, 5.8S, 5S | 47 |
| Mammalian mitochondria | | 55S | | |
| | Small subunit | 28S | 12S | 29 |
| | Large subunit | 39S | 16S | 47 |

TABLE 11.3 The Molecular Composition of Ribosomes from the Three Kingdoms

The two subunits are in contact with each other through a number of inter-subunit bridges (B1a-B8, Figure 11.14, left). rRNA, proteins as well as magnesium ions and water molecules mediate these bridging contacts. On the ribosome, there are three main sites for tRNA molecules, the A (aminoacyl), P (peptidyl) and E (exit) sites (Figure 11.14). An additional site, the T site, is the initial binding site where the tRNA binds in complex with elongation factor EF-Tu and GTP.

11.3.2 A Brief Summary of the Steps in Translation

Translation is normally described in four stages: initiation, elongation, termination and recycling (Figure 11.15). During initiation, the mRNA binds to the small subunit and the initiator tRNA molecule binds to the P site of the small subunit. Subsequently, the large subunit will associate. In bacteria, initiation is assisted and catalyzed by three initiation factors.

The elongation phase can be divided into three steps: decoding with the accommodation of aminoacyl-tRNA, peptidyl transfer and translocation. The polypeptide chain grows while the ribosome translates the mRNA and travels along the mRNA. At the stop codon, the polypeptide is released and subsequently all the components dissociate to be recycled. In most of these steps, different protein factors are involved.

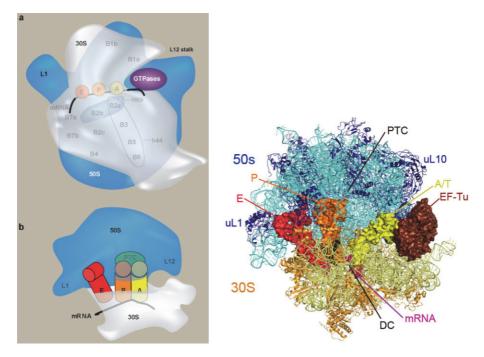


Fig. 11.14 ■ *Left top*: A schematic illustration of the organization of the bacterial ribosome. The large subunit (50S) is seen in the background and the small subunit (30S) is in the foreground. The functional sites are between the ribosomal subunits. The mRNA is bound around the neck of the small subunit, between the head and the body. Three sites for tRNA are shown, the A, P and E sites. The binding site for the translational factors, the trGTPases is also shown. A number of bridges (B1-B8) between the subunits are also illustrated. Two RNA helices are of special interest, h44 (30S) and H69 (50S). They make a functionally important interaction at bridge B2a that is at the decoding site (A site) for tRNA. Left bottom: The ribosome seen from top with the mRNA, the sites for tRNA and the peptidyl transfer center (PTC). (Reprinted with permission from Liljas A. (2006) Deepening ribosomal insights. ACS Chem. Biol 1: 567–569. Copyright ACS.) Right: The crystal structure of 70S ribosomes from T. thermophilus with tRNAs in the E and P sites and a tRNA bound to the hybrid A/T site in complex with EF-Tu. The RNA molecules of the large subunit are seen in light blue and the proteins are dark blue. The RNA of the small subunit is shown in yellow and the proteins are brown. The tRNA at the E, P and A/T sites are red, orange and yellow, and EF-Tu is in red. (Drawing made by Saraboji Kadhirvel, PDB: 2WRN and 2WRO.)

11.3.3 Structural Studies of the Ribosome

The ribosome is a challenging object for structural studies. The structures of complete ribosomes from two bacterial species (E. coli and Thermus thermophilus) have been characterized in many different functional stages. More recently, the crystal structures of eukaryotic ribosomes from humans, yeast and Tetrahymena thermophila have become available. In addition, mitochondrial ribosomes from humans and yeast have also been analyzed.

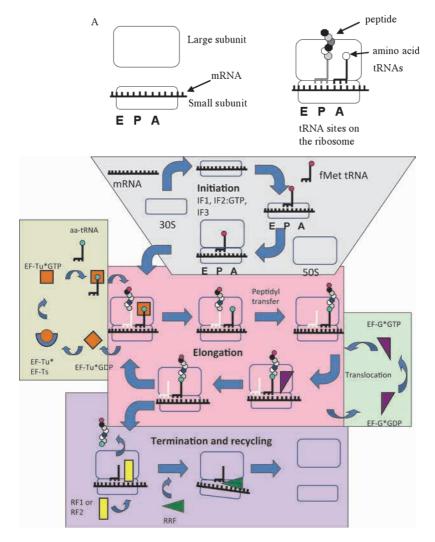


Fig. 11.15 ■ A schematic representation of the main steps of protein synthesis in bacteria.

Crystallography has been the main technique, but electron microscopy techniques have increasingly contributed essential insights into ribosome structure and function.

An important result is that the ribosome is flexible. The ribosomal subunits move with respect to one another during functional steps and domains of the subunits are also flexible. The main movement is a 6° anticlockwise rotation of the small subunit with regard to the large subunit associated with the binding of tRNAs to hybrid sites (A/P, P/E). The subunits return to their standard relative orientations along with movements of the tRNAs into their classical sites associated with hydrolysis of GTP by the trGTPase translation factors. Some of the inter-subunit contacts are changed during the translocation step.

11.3.3.1 Structure of the large subunit

The large subunit has the shape of a crown when viewed from the interface side, and is half-spherical when seen from the side. The three extensions are (from left to right): the uL1 stalk, the central protuberance (where the 5S rRNA is located) and the bL12 stalk. The two side protuberances have proteins as major parts and have significant flexibility related to their function.

The large subunit rRNA forms the core of the subunit. The secondary structure of the bacterial 23S rRNA is composed of approximately 100 double-stranded helices forming six domains. The domains are entangled with one another to give the core of the large subunit a stable structure. There are also many tertiary structure interactions.

The proteins are primarily located on the surface of the ribosome. Many of them have unusual structures (Figure 11.16), with globular parts on the surface of the ribosome and extensions that interact with the rRNA in the interior of the subunit. Some ribosomal proteins are completely extended, interacting with the rRNA. These extensions are important for the assembly and stability of the ribosome.

The main function of the large subunit is to catalyze the formation of the peptide bond. This is done in the peptidyl transfer center (PTC). The PTC is primarily composed of the 23S rRNA, but in bacterial ribosomes the N-terminal tail of protein bL27 is located in the PTC and known to be important for peptidyl transfer.

The growing or nascent polypeptide emerges through a tunnel through the large subunit (Figure 11.17). This exit tunnel begins at the peptidyl transfer center of the large subunit and ends at the outer surface of the large subunit. This 100 Å long tunnel is formed by rRNA and a few proteins.

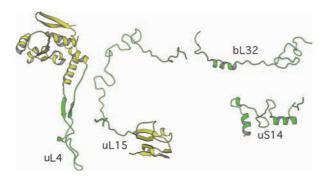


Fig. 11.16 ■ The structures of some ribosomal proteins illustrating their unusual conformations. A majority of ribosomal proteins have extended termini, long loops or domains separated by linkers of variable length. Parts of the proteins that are likely to fold only when bound to the ribosome are shown in green. These proteins are from the *T. thermophilus* ribosome (PDB: 2J00 and 2J01).

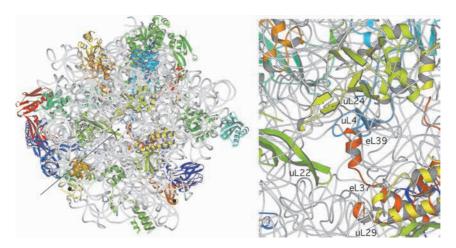


Fig. 11.17 ■ *Left*: A view of the large archaeal ribosomal subunit looking down the peptide exit tunnel (arrow). The RNA is shown as a coil. *Right*: A magnification of the peptide exit tunnel. Extended parts of proteins uL4 and uL22 are important parts of the tunnel (PDB: 1FFK).

11.3.3.2 The small subunit

Like the large subunit, the core of the small subunit is composed of rRNA. The four domains of the subunit are more clearly separated than the ones of the 50S subunit, leading to higher variability in their relative orientation. In Figure 11.18, the body of the subunit (red) is formed by the 5' third of the 16S rRNA. The platform (green) is the central part of the 16S rRNA and the head of the subunit (yellow) with the beak (*left*) is formed by the next segments. The 3'-end (blue) forms a long vertical helix (helix h44). The proteins are mainly on the surface and partly connecting and stabilizing contacts between the rRNA helices.

The small subunit binds the mRNA around its neck region between the head and the body of the subunit. The main function of the small subunit is to participate in the decoding of the mRNA. It also forms parts of the A, P and E sites. These sites have been defined with increasing clarity from biochemical experiments, crystallography and electron microscopy of the whole ribosome.

11.3.3.3 Eukaryotic ribosomes

The eukaryotic 80S ribosome and its 40S and 60S subunits have been studied by cryo-EM and crystallography. The main outlines of the ribosome are very similar to those of bacteria. However, as Table 11.3 shows, it has larger rRNA molecules and a greater number of ribosomal proteins (r-proteins). The extra proteins and the extension segments of the 18S rRNA are primarily located on the outer surface of the subunits, whereas the subunit

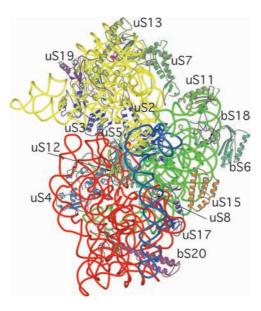


Fig. 11.18 ■ The small ribosomal subunit with the RNA domains illustrated from the 5′- to 3′-end by red, green, yellow and blue ribbons. h44 is seen as a blue helix extending from the lower right upto the junction between the body and the head of the subunit. Most of the proteins are also marked (PDB: 1FJG).

interface is more conserved in the three kingdoms. More than half of the conserved proteins have extra extensions in eukaryotes. Furthermore, the ribosomal proteins have more extensive contacts in eukaryotes than in bacteria.

In the 55S porcine and human mitochondrial ribosome, the rRNA is much reduced in size while there are a larger number of proteins. The comparison of these ribosomes with bacterial ones shows that the helices of the 12S and 16S rRNA have been shortened compared to bacterial ribosomes. The trimmed rRNA is covered with an outer layer of proteins. The interface between the subunits is more open. The additional proteins do not replace specific parts of the rRNA, but just substitute for the lost volume. The mitochondrial large subunit is larger than the bacterial one despite the reduction of the rRNA. In particular, the 5S rRNA is absent. However, the central protuberance, which is the location of the 5S rRNA, is double in size due to the binding of the bound proteins. The polypeptide exit tunnel is more open in mitochondrial ribosomes due to the reduced size of the large subunit rRNA.

11.4 Initiation

The initiation codon, AUG, also codes for methionine. To avoid initiation from any random AUG codon the correct initiator AUG codon has to be presented in the P site. This is

IF1 is a small universally conserved protein that binds to the decoding part of the 30S A site. This position of IF1 prevents the initiator tRNA from binding at the A site, allowing it to bind only at the P site. In addition, it participates in stabilizing the codon-anticodon interaction between mRNA and initiator tRNA.

IF3 binds to the small subunit of the ribosome, and prevents the small and large subunit from assembling prematurely, also guiding the initiator tRNA into the P site. IF3 has two well-separated domains. The N-terminal domain, IF3N, binds to the 30S part of the E site, while IF3C binds at the interface side of the platform. This binding site prevents H69 of the large subunit from interacting with the small subunit through inter-subunit bridge B2b.

IF2 belongs to the family of G-proteins and is one of the translational GTPases (trGTPases, Section 8.3.1). It catalyzes the binding of the small subunit to the large subunit to complete the initiation complex.

In eukaryotes, initiation is more complex than in prokaryotes and many more factors are involved. Two G-proteins are involved in initiation. The role of eIF2 is to bind the initiator tRNA to the ribosome. Another G-protein, eIF5B, is the homologue of the bacterial IF2 and catalyzes the joining of the ribosomal subunits. It has four domains, of which the N-terminal domain is the G domain, similar to Ras and other G-proteins. The first two domains are similar to the corresponding domains in EF-Tu.

11.5 Elongation

In bacterial elongation, two main protein factors catalyze the binding and translocation of the tRNAs (Table 11.4). They are called EF-Tu and EF-G. Both these proteins are G-proteins (Section 8.3.1), and normally they hydrolyze one GTP molecule each to GDP for each elongation cycle. In the elongation cycle, an aminoacyl-tRNA binds to the A site with the aid of EF-Tu in complex with GTP. After peptidyl transfer, EF-G catalyzes the translocation step in which the peptidyl tRNA in the A site will be moved into the P site and the deacylated tRNA will be moved from P to E site. In this process, the mRNA will also be moved forward to expose a new codon in the A site.

11.5.1 Elongation Factor EF-Tu

Elongation factor EF-Tu is responsible for delivering the aminoacylated tRNA to the ribosome. EF-Tu is a three-domain protein. The N-terminal domain is the G domain with a

TABLE 11.4 Elongation Factors in Bacteria and Eukaryotes. The Name in Parenthesis is the One Mostly Used for the Bacterial Proteins

| Protein | Function | | | |
|--------------------|---|--|--|--|
| EF1A (EF-Tu) | trGTPases involved in binding of aa-tRNA to ribosomes. | | | |
| (SelB) | The protein corresponding to EF-Tu for selenocysteine, a rare 21 st amino acid incorporated in some proteins using a unique tRNA. | | | |
| EF1B (EF-Ts) | Nucleotide exchange factor for EF1B/EF-Tu. The bacterial and eukaryotic proteins are unrelated. The eukaryotic eEF1B has two subunits, α and γ , where α is the active exchange factor and γ is very similar to the enzyme glutathione transferase. | | | |
| EF2 (EF-G) | trGTPase involved in the translocation of peptidyl-tRNA from A to P site or ribosome. | | | |
| EF3 | Probably facilitates the release of E site deacylated tRNA in eukaryotes. | | | |
| EF4 (Lep A) | trGTPase that acts as a putative back translocase. It appears to move the tRNAs and mRNA in the opposite direction of EF2 (EF-G). | | | |
| (Tet M, O, S etc.) | Tetracycline resistance factors. Close homologues of EF-G. | | | |

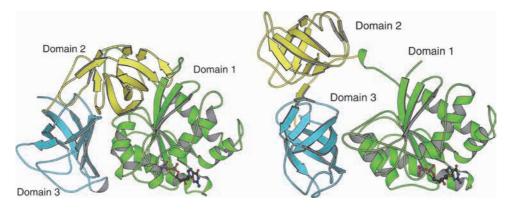


Fig. 11.19 ■ Schematic drawing of *T. aquaticus* EF-Tu in the GTP conformation (*left*, PDB: 1EFT) and the *E. coli* EF-Tu in the GDP conformation (*right*, PDB: 1TUI).

mainly parallel β -sheet with the same topology as other G-proteins (Figures 11.19 and 11.20). The other two domains (2 and 3) are both antiparallel β -barrels.

There is a large conformational difference between the GTP and the GDP forms. In the GTP form, the three domains are packed closely together, but in the GDP form, the two β-barrel domains move away from the G domain to generate a more open conformation (Figure 11.19) Some regions of the protein differ by 40 Å between the two conformations. Domains 2 and 3 retain their relative orientation in the two forms.

GTP is bound in the same way in EF-Tu as in other G-proteins. The P-loop binds the α - and β -phosphates, and a magnesium ion coordinates the β -phosphate in GDP and the

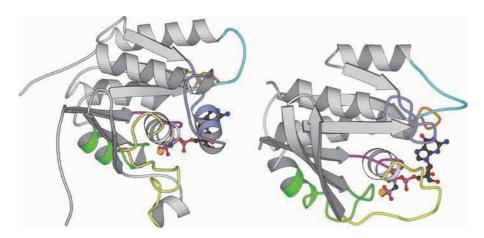


Fig. 11.20 ■ A comparison of the structure of the N-terminal domain 1 of EF-Tu (*left*, PDB: 1EFT) and the signaling protein Ras (right, PDB: 121P). The loops contacting the nucleotide are purple (G1, P-loop), yellow (G2, switch I), green (G3, switch II), turquoise (G4) and light blue (G5). A sixth loop that does not form any direct contact is shown in brown. The magnesium ion is orange.

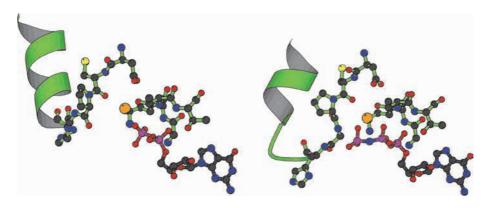


Fig. 11.21 ■ Detail of the binding of GDP/GTP showing the P-loop and switch II of EF-Tu. Left: The GDP conformation with GDPNP inserted (pale color), showing the close contact between the O83 carbonyl oxygen and the γ -phosphate. Right: The GTP conformation with the new position and orientation of the helix.

β- and γ-phosphates in GTP, as in Ras (Figure 11.20). The switch regions show different conformational changes compared to Ras and trimeric G-proteins (Section 14.4.2).

The GDP and GTP conformations of switches I and II of EF-Tu are different (Figure 11.21). In the GDP form, switch I is a β-ribbon, while in the GTP form it adopts a helical conformation. The helix of switch II in the GDP form is composed of residues 85 to 94. With GTP, the loop preceding this helix has to change conformation, since the carbonyl oxygen of residue 83 would be too close to the γ-phosphate. This peptide is therefore "flipped" in the GDP form, to allow the peptide nitrogen to hydrogen bond to the

phosphate. This has the effect of "dissolving" the helix. The loop is then extended to residue 88, and the helix (now between residues 89 and 96) has a different orientation. This helix is tightly associated with domain 3 in the GTP form. Thus, GTP alters the conformation of switch II to generate a surface that can associate with domains II and III, and activate the factor for tRNA binding. The translational GTPases have a histidine residue (His 84 in EF-Tu of T. thermophilus) in contrast to other GTPases, like Ras, that have a glutamine (Section 8.3). This residue is involved in GTP hydrolysis by placing a water molecule at the γ -phosphate for the hydrolysis.

11.5.1.1 Binding to tRNA — the ternary complex

The activated form of EF-Tu (with GTP) binds tRNA in a very different way from the binding to the aminoacyl tRNA synthetases. Since EF-Tu can bind all the different tRNAs, it has to recognize common features of the molecule. In the complex, all three domains of EF-Tu interact with the tRNA molecule but only the acceptor end, with its conserved CCA sequence, is involved in the contact. The anticodon stem and loop are pointing away from EF-Tu (Figure 11.22). The discrimination against uncharged tRNAs is very high. This may be somewhat surprising since the presence of a single amino acid does not change the surface of the tRNA significantly. However, the specificity for a charged tRNA is achieved through interactions with the amino group and the ester bond that connects the amino acid to the ribose of 3' nucleotide. The 3'-end with the amino acid is bound at the interface of EF-Tu domains 1 and 2. The 5'-end of the tRNA is bound at the intersection of all three domains.

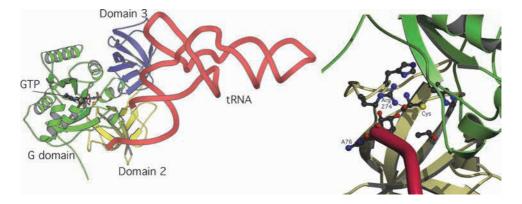


Fig. 11.22 ■ Left: The complex between EF-Tu and Cys tRNA. Right: Details of the interaction of the 3' (CCA)-end (A76) and the amino acid (Cys) of the tRNA. The guanidinium group of an arginine (Arg274) hydrogen bonds to the carbonyl oxygen of the bound amino acid while other side chains from both domain 1 (the G domain) and domain 2 form a cage around it (PDB: 1B23).

11.5.2 tRNA Binding and Decoding

In the first step of elongation on the ribosome, the anticodon of the ternary complex of aminoacyl-tRNA with EF-Tu-GTP is tested against the codon in the A site. The structure of the ternary complex bound to the ribosome has also been studied. The antibiotic kirromycin was used to keep EF-Tu from dissociating from the ribosome. In this case, the GTP is hydrolyzed to GDP, but EF-Tu cannot be released because the kirromycin prevents the factor from relaxing to the GDP conformation. Also, the non-hydrolyzable GTP-analogue GDPCP has been used for binding the factor in an activated state.

The binding of the ternary complex to the factor-binding site of the ribosome is primarily due to EF-Tu. The interaction surface of the tRNA with the ribosome is limited. Since EF-Tu binds to the factor-binding site of the ribosome and to the acceptor end of the tRNA, the amino acid cannot arrive at the peptidyl transfer site. The initial binding of the aminoacyl-tRNA is thus not to the A site but in the so-called T state (Figure 11.23). The anticodon stem and loop (ASL) can interact with the codon in the A site only by generating a bend in the tRNA. This bend occurs between the ASL and D stems. In this bent conformation, the tRNA binds to the A/T site (Figure 11.24).

In this step, which is called initial selection, the ribosome participates to identify cognate from non-cognate interactions. Correct Watson-Crick base-pairing of the first and second base pairs of the codon-anticodon is identified through hydrogen bonding by A1492 and A1493 of helix h44 as well as G530 to the base pairs formed by the codon and anticodon (Figure 11.24). This stabilizes the cognate codon-anticodon interaction.

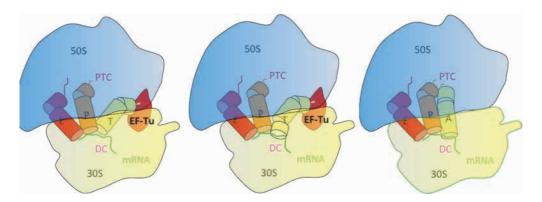


Fig. 11.23 ■ A simplified illustration of the binding of tRNA to the ribosome. *Left*: The ternary complex of EF-Tu, tRNA and GTP has bound to the T site. Middle: To interact with the mRNA a bend develops in the tRNA. Right: If the anticodon matches the codon EF-Tu will dissociate after GTP hydrolysis. Subsequently, the tRNA can accommodate into the A site.

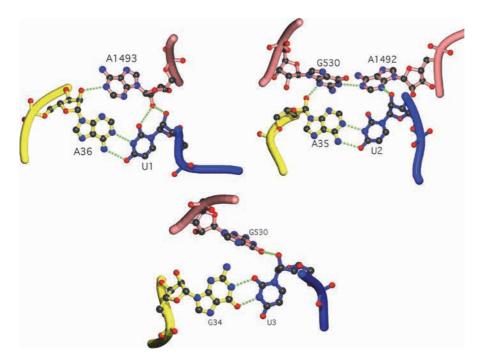


Fig. 11.24 ■ The fidelity of translation is based on the correctness of the codon-anticodon interaction. The ribosome participates in this activity through the rRNA of the small subunit. The mRNA is blue; the anticodon of the tRNA is in yellow and the ribosomal 16S rRNA is in pink. The Watson-Crick base paring of the two first base pairs in the codon-anticodon interaction is checked by A1493, A1492 and G530 of the 16S rRNA (top left and right, PDB: 2J00). Only the correct base pairs between codon and anticodon will allow the network of hydrogen bonds to form. For base pairing of other types than the Watson-Crick type, such as a U-G base pair in the first position, A1493 cannot interact properly with the codon ribose. The tRNA will, in such cases, not form a stable interaction with the ribosome and fall off. The interactions of the third (wobble) base pair (here a GU pair) are less strict (bottom, PDB: 1IBL).

11.5.2.1 GTP hydrolysis

A region of the 23S rRNA has been identified to be involved in inducing GTP hydrolysis. The region is called the sarcin-ricin loop (SRL; see Section 5.3.9) after two inhibitory enzymes that can covalently modify this region of the 23S rRNA. The modifications lead to the loss of proper functions of trGTPases. The functional component corresponding to a GAP (see Section 8.3.3) for other GTPases is the phosphate of A2662, not a protein. Through a conformational change His 84 interacts with this phosphate, which leads to a movement of the water molecule close to the γ -phosphate of the GTP. The histidine is positively charged due to its closeness to two negatively charged phosphates and can only be a donor of a hydrogen bond to the water molecule. The water molecule in turn

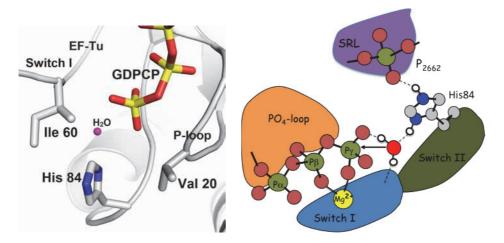


Fig. 11.25 • When EF-Tu is bound to the ribosome with a tRNA that is cognate with the codon in the decoding site and with a GTP analog, His84 is stabilized in a new position. *Left*: His 84 interacting with A2662 of the 23S rRNA moves between two hydrophobic residues and comes into contact with the water molecule, which is pushed against the γ-phosphate (PDB: 2xqd and 2xqe. (Kindly provided by Saraboji Kadhirvel.) *Right*: The positively charged histidine makes the water molecule donate hydrogen bonds to a carbonyl group in switch I and to an oxygen of the γ-phosphate. This leads to the activation of the water molecule by transferring a proton to the γ-phosphate and an attack of the generated hydroxyl on the GTP molecule, which is converted to GDP.

donates a proton to the γ -phosphate, which becomes attacked by the resulting hydroxyl leading to GTP hydrolysis (Figure 11.25). All translational GTPases have a histidine, rather than the normal glutamine in other GTPases.

Once the GTP of EF-Tu is hydrolyzed, the factor can dissociate from the ribosome and the tRNA can accommodate into the A site (Figure 11.23). In this accommodation step, there is a second chance for a non-cognate tRNA to fall off from the ribosome through so-called proofreading.

11.5.2.2 Elongation factor Ts, the G-nucleotide exchange factor

Elongation factor Ts (EF-Ts) is the G-nucleotide exchange factor (GEF) for EF-Tu. This protein is less conserved than EF-Tu. Crystal structures of EF-Tu·EF-Ts complexes are available. Three changes in the structure of EF-Tu explain how EF-Ts can act as a nucleotide exchange factor. In the complex, the B helix of switch II is moved, which leads to a loss of ligands to the magnesium ion. Another change is that Phe82 of EF-Ts is inserted into a pocket of EF-Tu, indirectly causing a displacement of the P-loop. The lysyl residue of the P-loop (K24), instead of binding to the β - and γ -phosphates of the nucleotide, now interacts with D81 of switch II (Figure 11.26). In addition, a flip of the peptide bond between residues 20

Fig. 11.26 ■ *Left:* A simplified illustration of some key interactions between EF-Tu and GTP. *Right:* The complex between EF-Tu·EF-Ts and GDP. EF-Ts induces Asp81 to interact with the P-loop lysine, K24. Both the binding of the magnesium and the nucleotide will be much weaker leading to nucleotide exchange.

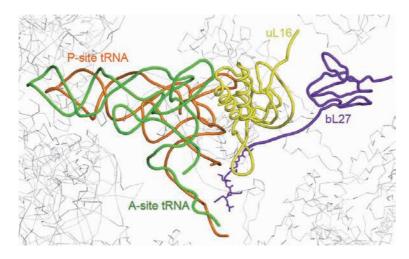


Fig. 11.27 ■ The A site (green) and P site (yellow) tRNAs when bound to the ribosome are close to two ribosomal proteins bL27 and uL16 which stabilize the position of the acceptor ends but do not directly participate in the peptidyl transfer activity (PDB: 2WDN and 2WDG. (Drawing made by Saraboji Kadhirvel.)

and 21 has the effect that the peptide nitrogen (N21), which binds the β -phosphate of GDP, is replaced by the carbonyl oxygen O20, which will repel the nucleotide.

11.5.3 Peptidyl Transfer

Peptidyl transfer does not require a catalyzing elongation factor. When the aminoacyl-tRNA is accommodated into the A site, the aminoacyl moiety is placed in the peptidyl transfer center (PTC) next to the nascent polypeptide that trails into the peptide exit channel of the large subunit (Figure 11.17). The PTC is mainly composed of the 23S rRNA. It has been claimed that bacterial ribosomes are ribozymes. However, the N-terminus of a protein, bL27, is located close to the ester bond between the tRNA and the growing polypeptide in the P site and is found to be important for full activity (Figure 11.27). Its role

may primarily be involved with stabilizing the acceptor ends of the tRNAs rather than a direct catalytic role.

The main steps of peptidyl transfer are shown in Figure 11.28. The amino group of the aminoacyl residue on the A site tRNA is bound by hydrogen bonds to A2451 and the 2'OH of A76 of the P site tRNA. This gives it the orientation and activation needed to attack the carboxyl carbon of the growing polypeptide in the P site that leads to peptidyl transfer.

During the steps of peptidyl transfer the nascent peptide is located in the peptidyl exit channel and remains in essentially the same place, but after the reaction it is connected to the A site tRNA. This places the tRNA again in a hybrid site, the A/P site (Figure 11.28). However, this requires that the P site tRNA has first moved into a P/E site. The main change is the 180° reorientation of the single-stranded CCA end of the A site tRNA to base pair with the loop in the P site part of the PTC. This 180° rotation of the CCA end of the

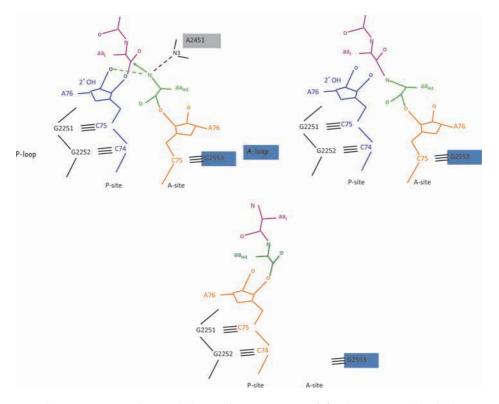


Fig. 11.28 ■ The main steps of peptidyl transfer in PTC. *Top left*: The CCA ends of the A- and P site tRNAs form base pairs to two specific loops of the 23S rRNA. The nascent polypeptide is located in the peptidyl exit channel and the amino group of the amino acyl moiety of the tRNA in the A site is properly oriented through hydrogen bonds to A2451 and the 2′OH of the terminal A of the P site tRNA. *Top right*: Peptidyl transfer has occurred. *Bottom*: The CCA end of the newly formed peptidyl tRNA replaces the deacylated tRNA in being bound to the P-loop and becomes bound to both A and P sites, thereby being bound to a hybrid A/P site.

peptidyl tRNA is facilitated by the two-fold symmetry of PTC. In the A site, C75 of the tRNA base pairs with G2553 of a loop of the 23S rRNA. After rotating into the P site, it base pairs with G2251 of a symmetry related loop. In addition, C74 base pairs with G2252 of the P-loop (Figure 11.28). The two-fold symmetry of PTC is due to 110 nucleotides in the A site that have a two-fold relation to 110 nucleotides in the P site.

11.5.4 Elongation Factor G

Elongation factor G (EF-G; EF2 in eukaryotes) catalyzes the translocation of the deacylated tRNA from P to E site and the peptidyl tRNA from A to P site on the ribosome. At the same time, the mRNA is moved to expose a new codon in the A site.

EF-G has five domains and is an elongated molecule (Figure 11.29). The N-terminal domain is the G domain. It has the same topology as other G-proteins. A subdomain (G'), consisting of an antiparallel sheet, is inserted before the last β -strand of the G domain. The second domain is an antiparallel β -barrel related to domain II in all trGTPases

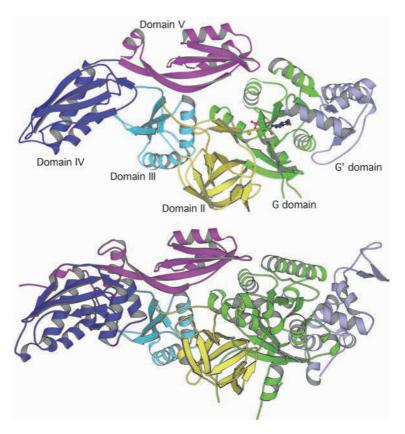


Fig. 11.29 ■ Schematic drawing of EF-G from *T. thermophilus* (*top*, PDB: 1FNM) and yeast EF-2 (*bottom*, PDB: 1N0V).

operating on the ribosome, for example, IF2 and EF-Tu. The other three domains are two-layer structures made up of antiparallel β -sheets with helices on one side. Domains III and V have the same topology as ribosomal protein S6, and many RNA-binding proteins. The EF-G molecule mimic the EF-Tu-tRNA complex (compare Figures 11.29 and 11.22). Domains III, IV and V mimic a tRNA molecule and EF-G binds to the ribosome just like the ternary complex of EF-Tu with tRNA.

EF-G has no nucleotide exchange factor. Its affinity for nucleotides is lower than that of EF-Tu and the exchange occurs spontaneously. In EF-Tu, the P-loop lysine (K24) and Asp81 from switch II stabilize the binding of GDP together with the magnesium ion that interacts with the phosphate. When EF-Tu interacts with its GEF, EF-Ts, the lysine and the aspartic acid instead are forced to interact with each other (Figure 11.26). This is the normal situation for EF-G. K25 interacts with T84 of switch II (Figure 11.31). These residues may interact respectively with the phosphates and magnesium when EF-G is bound to the ribosome but not on EF-G in solution. Thus, both the metal ion and the nucleotide are more loosely bound and GDP can be exchanged for GTP.

11.5.5 Translocation

The final step of the elongation cycle is translocation catalyzed by EF-G. As discussed above, the acceptor end of the peptidyl tRNA in the A site spontaneously translocates by an 180° rotation into the P site generating a hybrid A/P state for the peptidyl tRNA. However, the main part of the tRNA remains in the A site. When EF-G binds to the ribosome it undergoes a conformational change by which it reaches into the decoding part of the A site and pushes the peptidyl-tRNA into the P site (Figure 11.30). EF-G·GTP, with its mimicry of the ternary complex aa-tRNA·EF-Tu·GTP, binds to the same site of the ribosome (compare Figures 11.23 and 11.30).

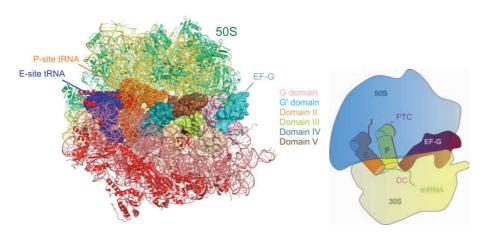


Fig. 11.30 ■ The structure of EF-G bound to the ribosome (PDB: 2WRI and 2WRJ). The binding is very similar to the binding of EF-Tu with a tRNA (Figure 11.23). (Drawing made by Saraboji Kadhirvel.)

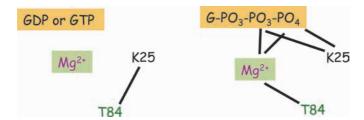


Fig. 11.31 ■ *Left:* EF-G has a conformation that allows spontaneous nucleotide exchange of GDP. The P-loop lysine normally interacts with T84 in switch II. *Right:* On the ribosome GTP is more firmly bound. This may be due to interactions of the lysyl and threonyl residues with phosphates and the magnesium ion, respectively.

EF-G binds a GTP molecule that is needed for its activity. The GTPase activity is induced in the same manner as for EF-Tu (Section 11.4.2.1). Probably, this GTP is hydrolyzed before translocation. If it would occur after translocation, EF-G behaves like a classical G-protein and a molecular switch. However, since it occurs before translocation, EF-G is more similar to a motor protein consuming GTP to do the work.

11.6 Peptide Release and Ribosome Recycling

11.6.1 Release Factors

There are three major release factors in bacteria, RF1, RF2 and RF3 (Table 11.5). RF1 and RF2 recognize the three stop codons (UAA, UAG and UGA) and promote the hydrolysis of the peptide from the tRNA at the P site. RF3 is a trGTPase that catalyzes the release of RF1 and RF2 from the ribosome. In eukaryotes, a single factor (eRF1) performs the function of RF1 and RF2, recognizing the stop codons, and eRF3 corresponds to RF3.

The eukaryotic release factor eRF1 is responsible for recognition of all three stop codons and stimulates the hydrolysis of the peptide. Since the mRNA and the peptide are at opposite ends of the tRNA in the P site, interactions at both sites require a molecule with similar dimensions and possibly with the same shape as tRNA. Human eRF1 has an elongated structure with three domains (Figure 11.32). A sequence motif found in eukaryotic as well as bacterial release factors, GGQ, is found at one extreme end of domain 2. This motif is important for the hydrolytic activity, performed by the peptidyl transfer site of the 23S rRNA in the ribosome. The other extreme end of the molecule is formed by two helices. Mutational studies suggest that this end interacts with the stop codon of the mRNA. Most RNA-binding proteins are based on antiparallel β -sheets, and this is thus an unusual conformation for binding to rRNA. To some extent, eRF1 has the size and shape of a tRNA molecule. It could bind to the stop codon at the A site and interact with the tRNA with the peptide at the P site in the same way

| | | | <u> </u> |
|----------|--------------------|------------|--|
| Bacteria | Stop Codons | Eukaryotes | Function |
| RF1 | UAA, UAG | eRF1 | Stop codon recognition and peptide hydrolysis. |
| RF2 | UAA, UGA | eRF1 | Stop codon recognition and peptide hydrolysis. |
| RF3 | | eRF3 | trGTPase, recycling the other release factors. |

TABLE 11.5 Release Factors in Prokarvotes and Eukarvotes

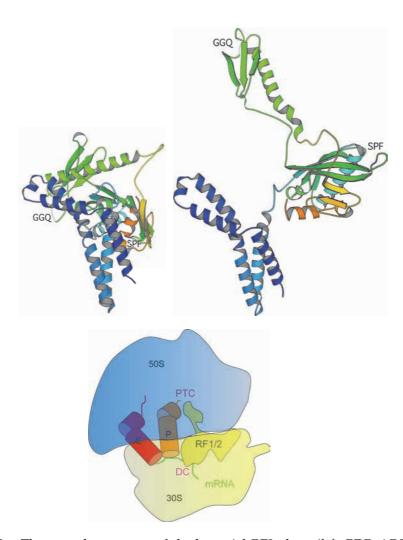


Fig. 11.32 ■ Top: The crystal structures of the bacterial RF2 alone (left, PDB: 1GQE) and on the ribosome (right, PDB: 1MI6). The distance between the motif involved in hydrolysis, GGQ and the motif that is at the decoding site, SPF is very different in the isolated state and on the ribosome. Bottom: The bacterial release factors bound to the ribosome.

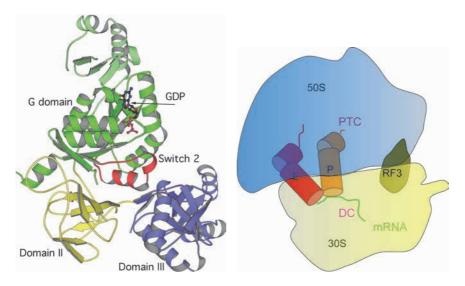


Fig. 11.33 ■ The structure of RF3 from bacteria (PDB: 3VQT) and its binding site on the ribosome.

as the A site tRNA during elongation. Thus, this could be yet another example of tRNA mimicry.

Bacterial release factors RF1 and RF2 perform the same function as eRF1. The bacterial release factors contain the same GGQ motif that is involved in peptide hydrolysis and release. Otherwise, the bacterial release factors have no sequence or structural similarity to the eukaryotic release factor. The distance between the GGQ motif and the tripeptide identified to be related to decoding the stop signal is no more than 23 Å in a crystal structure of the isolated RF2. This is much shorter than the distance between the decoding site of the small subunit and PTC of the large subunit. However, the conformation of this protein is very different when bound to the ribosome (Figure 11.32). This large conformational change has been confirmed by several independent structure determinations. Its functional importance remains to be explained.

The role of RF3 is to release RF1 or RF2 from the ribosome, RF3 is a GTPase with a structure much like EF-Tu (Figure 11.19). Its binding site partly overlaps with the site for RF1/2 (Figure 11.33).

11.6.2 The Ribosome Recycling Factor

After the peptide release, the mRNA and a deacylated tRNA remain bound to the ribosome. The dissociation of this complex is due to another protein, the ribosome recycling factor (RRF). It consists of two domains. Domain I has three long, almost parallel helices. The other domain is formed by the sequence after the N-terminal long helix. The molecule is L-shaped, and its dimension and shape are very close to those of a tRNA molecule (Figure 11.34).

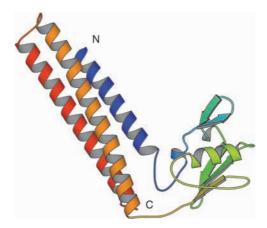


Fig. 11.34 ■ The ribosome recycling factor from *Thermotoga maritima* (PDB: 1DD5).

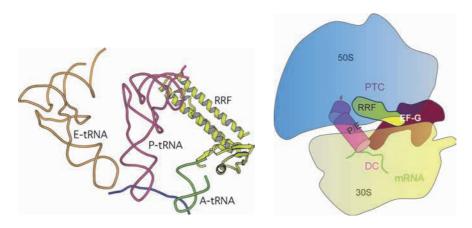


Fig. 11.35 ■ Left: RRF, when bound to the 70S ribosome, binds across parts of both A and P sites. Its binding moves the acceptor stem of the P site tRNA towards the E site Right: PDB: 2J00 and 2V46). RRF functions in conjunction with EF-G to split the ribosome into subunits.

Multiple structures of RRF have been determined. This has illustrated that the hinge between the two domains is quite flexible. RRF is another example of tRNA mimicry. However, the protein does not bind to a single tRNA site on the ribosome. It rather binds across acceptor stem parts of the A and P sites (Figure 11.35).

RRF requires the assistance of EF-G and GTP to perform its activity of splitting the ribosome into subunits. This induces a conformational change of the two domains of RRF that breaks essential subunit contacts. The recycling step is the fourth and last step in the translation cycle. The subsequent binding of IF3 prevents the premature reunion of the subunits.

Further Reading (Sections 11.1–11.2)

Original Articles

Crick FHC. (1958) On protein synthesis. Symp Soc Exp Biol 12: 138–163.

Hoagland MB, Zamecnic P, Stephenson ML. (1957) Intermediate reactions in protein biosynthesis. Biochim Biophys Acta 24: 2015-2016.

Holley RW, Apgar J, Everett GA et al. (1965) Structure of a ribonucleic acid. Science 147: 1462–1465.

Reviews

- Giege R, Sissler M, Florenz C. (1998) Universal rules and idiosynchratic features in tRNA identity. Nucl Acids Res 26: 5017-5035.
- Fukai S, Nureki O, Sekine S, et al. (2000) Structural basis for double-sieve discrimination of L-valine from L-isoleucine and L-threonine by the complex of tRNAVal and Valyl-tRNA Synthetase. Cell 103: 793-803.
- Vasileva IA, Moor NA. (2007) Interaction of aminoacyl tRNA synthetases with tRNA: General principles and distinguishing characteristics of the high molecular weight substrate recognition. Biochemistry (Moscow) 72: 306–324.

Further Reading (Sections 11.3-11.6)

Original Articles

- Carvalho ATP, Szeler K, Vavitsas K, et al. (2015) Modeling the mechanisms of biological GTP hydrolysis. *ABB* **582**: 80–90.
- Ban N, Nissen P, Hansen J, et al. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**: 905–920.
- Liljas A, Åqvist J, Ehrenberg M. (2011) Comment on "The mechanism for activation of GTP hydrolysis on the ribosome". Science 333: 37a.
- Selmer M, Dunham CM, Murphy FV, et al. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. Science 313: 1935–1942.
- Schmeing TM, Huang KS, Strobel SA, Steitz TA. (2005) An induced-fit mechanism to promote peptide bond formation and exclude hydrolysis of peptidyl-tRNA. Nature 438: 520-524.
- Voorhees RM, Schmeing M, Kelley AC, Ramakrishnan V. (2010) The mechanism for activation of GTP hydrolysis on the ribosome. *Science* **330**: 835–838.
- Wimberly BT, Brodersen DE, Clemons WM, et al. (2000) Structure of the 30S ribosomal subunit. Nature 407: 327-339.

Yusupova GZ, Yusupov MM, Cate JH, Noller HF. (2001) The path of messenger RNA through the ribosome. *Cell* **106**: 233–241.

Zhou J, Lancaster L, Donohue JP, Noller HF. (2013) Crystal structures of EF-G-ribosome complexes trapped in intermediate states of translocation. *Science* **340**: 1236086, 1–9.

Reviews

Liljas A, Ehrenberg M. (2013) Structural Aspects of Protein Synthesis, 2nd edn. World Scientific, Singapore.

Moore PB. (2009) The ribosome returned. J. Biol. 8(8), 1–10.

Protein Folding and Degradation

12.1 Protein Folding

Protein synthesis is a central theme in biological sciences. However, proteins would not function unless they fold correctly into their specific three-dimensional structure. Half a century ago, experiments by Anfinsen showed that protein folding could be a spontaneous process since a functional staphylococcal nuclease could be refolded from denatured fragments of the protein. The amino acid sequence determines the fold of the protein in the *folding process*. An analysis of the amino acid sequence could thus presumably identify structural features and maybe the fold of the protein. This is a fundamental problem that has still not been satisfactorily solved. It seemed unlikely that proteins would assist in the folding process. How could there be molds for such a variety of folds, and in turn, how were the molds folded?

Another aspect was the rate of folding. Levinthal pointed out that for a polypeptide to search through all of the possible conformations would take longer than the age of the universe. Furthermore, the total protein concentration is about 300 mg/ml in the cell and the environment is hostile due to the proteolytic enzymes. If the protein is not properly folded it may aggregate or be degraded. However, experimentally *in vitro* it has been established that the folding of a single-domain protein from a denatured state to a functionally active conformation can take in the order of milliseconds to seconds. Evidently, folding must follow some path that avoids the Levinthal scenario. Local structures along the polypeptide are likely to initiate the folding.

12.1.1 Spontaneous and Assisted Protein Folding

The spontaneous folding of proteins is described in Section 3.1.1.1. Many proteins fold spontaneously when they are synthesized. However, the conditions in the cell can be

such that some proteins may lose or never attain its native structure. Thus, there are conditions, e.g. elevated temperatures, where proteins need assistance to arrive at or regain their functional structure. Proteins that are not properly folded may aggregate into damaging amyloid structures (Section 3.3.3), which can be deleterious for the organism. For example, in overexpression of proteins, inclusion bodies are frequently formed. However, nature has developed a number of rescue mechanisms to assist proteins in need of help with folding. As a collective description, these systems provide an *assisted self-assembly* for protein structures. The mechanisms are highly conserved throughout evolution, indicating their vital role in all living systems.

The first insight into assisted folding processes came with the discovery of peptidyl proline *cis-trans* isomerases (PPIases) and protein disulfide isomerases (PDIs). The folding rate of some proline-containing proteins was greatly accelerated by adding proteins that specifically assisted the conversion between the *trans* and *cis* conformations of the main chain at the proline residue.

In parallel, many proteins induced by thermal stress in the cells were identified. They were called heat shock proteins (Hsp). Some of them rescue thermally denatured proteins. Proteins of various sizes were observed and homologies between proteins from different species were found. The idea arose that perhaps these proteins not only had a function under stress conditions but also under normal conditions in the cell.

12.1.1.1 Proline cis-trans isomerases

For proline residues, both the cis- and trans-peptide conformations are accessible since the peptide torsion angle, ω can adopt both 0° and 180° (Section 2.2.4). Some prolines need the cis conformation for the rest of the protein to fold properly, and proline isomerization can reduce the rate of protein folding significantly. The first proline cis-trans isomerase (PPIase) was isolated in 1984. At the same time, the protein cyclophilin (CypA), a receptor of the immunosuppressive drugs cyclosporine A was identified. The receptor (FKBP) for another immunosuppressive drug, FK506 was also isolated. Some years later, it was realized that these receptors were identical to the PPIases and the rate of acceleration by the enzyme can be $10^3 - 10^6$ compared with the spontaneous process. We now know that there are at least three families of PPIases: the CyPs, the FKBPs and the parvulins. These three families are represented in all forms of life. The family size varies with the complexity of the proteome. The catalytic domains of the three families are structurally different and so is the substrate specificity. The proper folding by PPIases of substrate proteins has distinct regulatory cellular roles. Thus, they regulate a wide range of physiology such as the activity on all levels of transcription, redox systems, heat shock response and photosynthesis. The PPIases can, in addition to the catalytic domain, have several extra domains important for the selectivity of substrates.

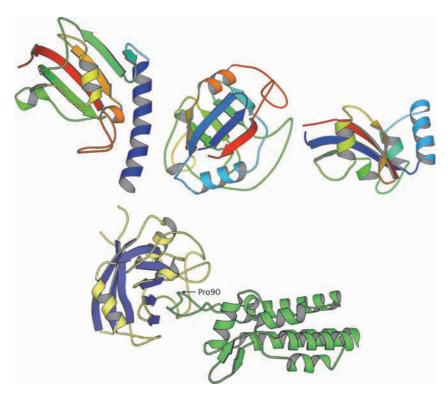


Fig. 12.1 ■ Top: The structures of the three main families of PPIases. Left: FKBP (PDB: 1Q6U), middle: cyclophilin (PDB: 2NUL), and right: parvulin (PDB: 1JNT). Bottom: The structure of cyclophilin A (left) in complex with the N-terminal domain of the HIV-1 CA (capsid) protein (right). The protruding loop of the virus protein contains a proline residue that undergoes trans-cis isomerization (PDB: 1M9C).

Crystal structures of PPIases with prolines in both conformations have been studied (Figures 12.1 and 12.2).

The total structural change caused by the isomerization is marginal, except that the carbonyl oxygen and the side chain preceding the proline undergo large shifts. As a consequence, their interactions with each other are different. Reducing the double bond character of the peptide induces catalysis.

PPIases are sometimes interacting with chaperones and can then be regarded as cochaperones (Section 12.1.2). In fact, some chaperones also contain a PPIase domain or function. One such example is the chaperone trigger factor that can interact with the growing polypeptide when it is emerging from the ribosome, having double functions as both a chaperone and a PPIase (Section 12.1.4.1).

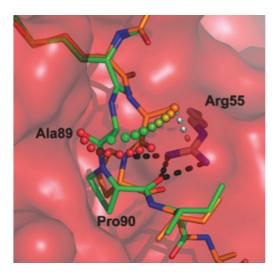


Fig. 12.2 ■ Close-up view of Pro90 in the HIV-1 CA protein bound to cyclophilin A. The cis conformation is shown in green and the *trans* in orange. The path of the conformational shift of Cβ and the carbonyl oxygen of Ala89 is illustrated by green or red spheres. The difference in conformation between the two structures is essentially limited to Ala89-Pro90. Arg55 of cyclophilin A limits the conformational changes of residues C-terminal of Pro90 by its hydrogen bond to the carbonyl oxygen of Pro90. In addition, a hydrogen bond between the arginine and the peptide nitrogen stabilizes the pyramidal sp³ hybridization of the proline nitrogen in the transition state and thereby also stabilize the single bond character of the peptide bond. This would favor catalysis. (Reprinted with permission from Howard BR, et al. (2003) Structural insight into the catalytic mechanism of cyclophilin A. Nat Struct Biol 10: 475-481. Copyright (2003) Nature.)

12.1.1.2 Protein disulfide isomerases (PDIs)

The protein disulfide isomerases (PDIs) were first isolated and characterized by Anfinsen in 1963. A protein with potential disulfide bonds needs to be oxidized for its sulfur atoms to form disulfide bridges. If there are several pairs of cysteines involved in disulfides, wrong connections can be made. In addition, oxidized cysteines can lead to aggregation. To avoid aggregation or incorrect folding, the incorrectly formed disulfides need to be reduced and new attempts made to reoxidize the protein into the right fold. The process is achieved somewhat differently in bacteria than eukaryotes.

DsbA is the protein initially oxidizing the disulfide-containing protein in bacteria while DsbC handles the isomerization (Figure 12.3). The bacterial enzymes are primarily found in the periplasm. In eukaryotes, only one enzyme is required for both oxidation and isomerization (Figure 12.4). The eukaryotic PDI is found in high concentrations in the endoplasmic reticulum (ER).

The oxidizing and isomerizing proteins contain one or several thioredoxin domains. In humans, there are four domains called a-b-b'-x-a', where x is a linker. The CxxC sequence motif (in humans CGHC) is the catalytic part of the a and a' domains, where the

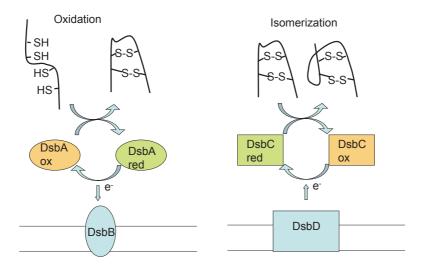


Fig. 12.3. ■ The oxidation and isomerization of disulfide-containing proteins in bacteria. The protein DsbA is responsible for the oxidation, whereas DsbB is a membrane protein that reoxidizes DsbA. The reduced form of DsbC catalyzes the isomerization. DsbC is kept reduced by DsbD.

two cysteines can be reduced or oxidized, forming a disulfide bond. The structure of full PDIs has been determined for a few species (Figure 12.4). A significant flexibility is observed in the U-shaped structure formed by the four domains but the -CGHC-active sites are seen facing each other. The x-linker is the major source of flexibility. The oxidized state is more open and exposes significant hydrophobic patches, which presents an inviting binding site for a client protein to become oxidized. Human PDI is re-oxidized by the proteins Ero1a and Ero1b, but structural data is lacking.

The activity of PDI is low in catalytic concentrations, but full and rapid functionality of substrate proteins is achieved at stoichiometric concentrations, which is the situation in ER. High restoration of substrate protein is gained even with PDIs where the active site cysteines have been blocked or mutated. Therefore, the PDIs can also be regarded as chaperones (see below). They can be claimed to be redox-dependent chaperones. The hydrophobic surfaces, particularly in domain b' are related to the chaperone activity.

12.1.2 Molecular Chaperones

The term "molecular chaperone" was first used in 1978 for a protein that prevented misassembly of nucleosomes. The connection to human situations was that a "chaperone" could prevent unsuitable interactions, in our case, unwanted aggregation with other proteins. Later, it was suggested that the growing group of heat shock proteins (Hsps) also had normal cellular functions since they were present even without heat stress. They were identified and classified as chaperones. Slowly there was an understanding that the chaperones not only prevented harmful aggregation, but that they also had the important role of assisting the folding of many proteins. Part of this function was to renature proteins

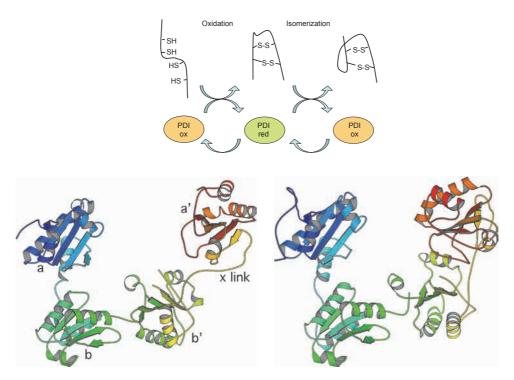


Fig. 12.4. Top: In eukaryotes, the same protein (PDI) is responsible for both oxidation and isomerization in the endoplasmic reticulum. PDI is reoxidized by the FAD-dependent ER oxidoreductin protein. Bottom: The structure of a human PDI in oxidized (PDB: 4EL1, left) and reduced (PDB: 4EKZ, right) states, illustrating some of the flexibility most clearly seen as the distance between the a and a' domains, which contain the active site components of the four thioredoxinlike domains.

that were unfolded due to heat or other denaturing conditions. This section will illuminate some basic structural insights about these chaperones.

The heat shock proteins (Table 12.1) can be divided into three main groups: the small heat shock proteins (sHsps), chaperones (usually monomeric proteins) and the chaperonins, which are large, hollow oligomeric structures. The two last groups of proteins use ATP to renature partly denatured proteins. The number after the "Hsp" indicates the approximate molecular mass in kDa of the chaperones.

12.1.2.1 Small heat shock proteins

In all domains of life, there is a large and diverse class of chaperones called small heat shock proteins (sHsps) of molecular masses between approximately 12 kDa and 40 kDa. Humans have around 10 different sHsps and C. elegans has 16. They are the first line of defense at stress conditions and bind proteins in partly folded conformations and prevent them from aggregation. sHsps are involved in many fundamental cellular processes and mutations in the sHsps are associated with human diseases including cataract, myopathy

Other Name **Functions** with Name Function sHsps Small heat shock Binding of non-native proteins Hsp40+Hsp70, Hsp100 proteins DnaK Hsp40 Hsp70 Refolding of non-native proteins Folding of newly synthesized proteins Hsp40 Binding of hydrophobic peptides. DnaJ Hsp70 Co-chaperone of Hsp70 **GrpE** Nucleotide exchange factor for Hsp70 Hsp70 Hsp110 Partly related to Hsp70 Hsp60 GroEL (chaperonin) Refolding of non-native proteins GroES Hsp10 **GroES** Co-chaperonin with GroEL GroEL Hsp90 Activation of regulatory and signaling proteins p23/Sba1 ClpA, ClpB ClpS Hsp100 Resolubilizes protein aggregates ClpS Hsp104 ClpB Hsp78 Mitochondrial ClpB homologue TF Trigger factor Assists folding of nascent chains Ribosomes

Some Representative Classes of Chaperones **TABLE 12.1**

and neuropathy. sHsps accumulate in cells where protein aggregation is frequent, as occurs in degenerative disorders such as Alzheimer's disease and Parkinson's disease.

The common feature of sHsps is a conserved α-crystallin domain (ACD) of about 80 amino acid residues. α-crystallin is a major constituent of the eye lens, which has a very high total protein concentration. α A- and α B-crystallins probably prevent proteins from aggregating in the lens leading to undesirable visual quality. The ACD can sometimes occur twice in an sHsp monomer. The ACD has an N-terminal extension of variable length and a short C-terminal extension (Figure 12.5a). α-crystallin has an immunoglobulin-like fold with a β -sheet sandwich. The two β -layers have three and four strands, respectively. The basic arrangement is a dimer formation where a long loop in one domain interacts with the edge of the sheet in the other (Figure 12.5b). In a second dimer variant, $\beta 7$ is fused with $\beta 6$ to form an antiparallel interaction with the same structure of another monomer (Figure 12.5c).

These crystallins regularly form large and dynamic oligomers with a hollow inside. The number of subunits of these oligomeric states varies (Figure 12.5d). The flexible terminal regions influence the size of the oligomers and are essential for the chaperone activity. The sHsps can form large complexes with non-native proteins. They can also become parts of inclusion bodies, which are large intracellular precipitates that occasionally form when large amounts of proteins are synthesized in expression systems.

The sHsps can be activated by heat. In eukaryotes, phosphorylation of the N-terminal extension also leads to activation. In the activated state, the small oligomeric forms of

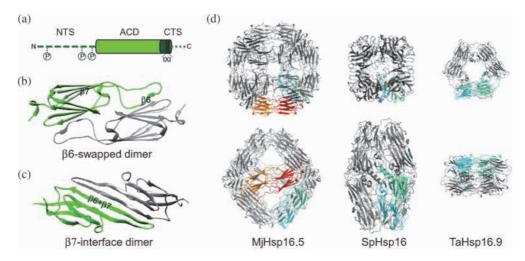


Fig. 12.5 • (a) The domain organization of sHsps. (b) A β6 swapped dimer of M. jannaschii Hsp16.5 (PDB: 1SHS). (c) A β7-interface dimer of human α B-crystallin (PDB: 2KLR). (d) 24-mer structure of M. jannaschii Hsp16.5; 16-mer structure of S. pombe Shsp16 and 12-mer structure of Tritium aestivum Hsp16.9. Dimers are highlighted in green/cyan and yellow/red. (Reproduced with permission from Haslbeck Vierling M,E. (2015) A first line of stress defense: Small heat shock proteins and their function in protein homeostasis. J Mol Biol 427: 1537–1548. Copyright (2015) Elsevier.)

sHsps increase to be able to bind client proteins. Subsequently, larger aggregates build up. The sHsps bind tightly to the non-native proteins and prevent them from aggregation or getting degraded. To release their substrates, they generally require assistance. This is provided by the larger chaperones that require ATP for the refolding reaction (Figure 12.6).

12.1.2.2 Chaperones

Many chaperones (Table 12.1) are monomeric or dimeric. The chaperone families Hsp70 and Hsp110 are monomers, whereas Hsp40 and Hsp90 are dimers. Hsp100, on the other hand, belongs to the AAA+ (Section 8.3.1) superfamily of proteins, which are hexameric aggregates. The binding and release of partly unfolded proteins by chaperones depends on the binding and hydrolysis of ATP by the chaperones.

12.1.2.3 Hsp70 and Hsp40

Many chaperones have co-chaperones, e.g. Hsp70 (DnaK in *E. coli*) has a co-chaperone Hsp40 (DnaJ in *E. coli*). This group of proteins is universal in bacteria and eukaryotes. While *E. coli* has 6 different types, yeast has 22 and humans 47 Hsp40s. There are two

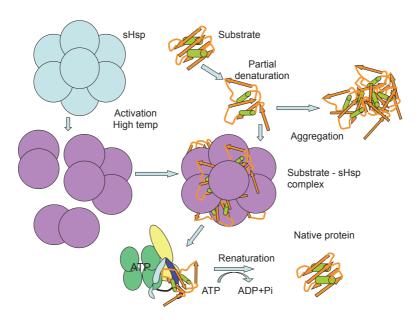


Fig. 12.6 ■ sHsps are built of dimeric units that can assemble into larger oligomeric aggregates. When activated by stress conditions, low molecular states are induced to allow proteins of nonnative conformation to bind and bigger aggregates are formed. This prevents proteins from forming irreversibly precipitated aggregates or getting proteolytically degraded. The complex between the sHsp oligomer and the substrate protein can be disassembled with the aid of monomeric chaperones like Hsp40, Hsp70 and Hsp100.

types of Hsp40 (1 and 2), but both function as dimers and have a large cleft between their two monomers. Hsp40 binds the unfolded substrate protein and prevents it from aggregating. Hsp70 interacts with Hsp40 and assists the unfolded protein to fold. The N-terminal domain of Hsp40 is called the J domain and can stimulate the ATPase activity of Hsp70 and regulate its binding of substrate proteins. The C-terminal region is composed of two or three domains. The peptide substrate binds as an antiparallel β -strand to a β-sheet in domain I (Figure 12.7).

Hsp70 is the ATP binding and hydrolyzing component of the couple. It has an N-terminal ATP-binding domain (NBD) with two lobes and it is composed of four subdomains (IA, IB, IIA and IIB) with an actin-like fold. The C-terminal substrate-binding domain (SBD) is composed of two subdomains, SBDα and SBDβ, where SBDβ binds the substrate and the former acts like a lid. The linker between NBD and SBD is flexible (Figure 12.8). Hsp70 can interact with a nucleotide exchange protein called GrpE. The binding and release of substrate proteins or peptides is regulated by the binding and hydrolysis of ATP.

Hsp70 binds to extended hydrophobic peptide segments of 5-7 residues flanked by positively charged residues. In the binding and release of the peptides, SBD is controlled by allosteric interactions between the ATP binding to NBD and hydrolysis. When ATP

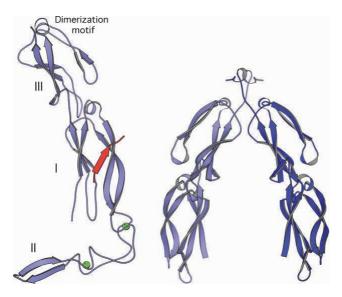


Fig. 12.7 ■ Left: A monomer of the peptide-binding C-terminal fragment of Hsp40 type 1 (Ydj1 from yeast, PDB: 1NLT)). The three domains I, II and III are marked. Two zinc-finger motifs and zinc ions (in green) are characteristic for Hsp40 type 1. The position for a bound substrate peptide is shown in red. Right: A dimer of type 2 (Sis1 from yeast, PDB: 1C3G). This protein lacks domain II.

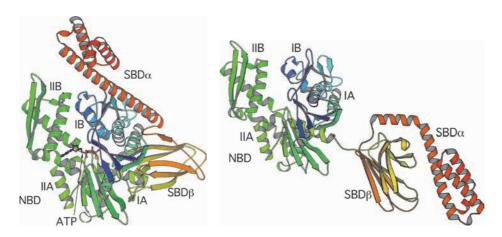


Fig. 12.8 ■ Two states of Hsp70 with ATP (left, low affinity state) and ADP (right, high affinity state). The nucleotide-binding subdomains IA, IB, IIA and IIB are shown in sequential coloring starting with blue at the N-terminus. ATP binds at the bottom of a cleft between subdomains IB and IIB. The substrate-binding subdomains SBDα (red) and in particular SBDβ (yellow) bind a peptide in the ADP state (PDB: 4B9Q and 2KHO). In the low affinity state, the ATP-binding site of NBD closes, opening a groove on the opposite side of NBD where the flexible linker between NBD and SBD can bind. This makes the substrate-binding site less accessible. In the high affinity state (with ADP), the nucleotide-binding site is open, which closes the groove for the linker making SBD able to interact with substrates.

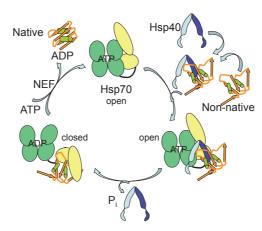


Fig. 12.9 ■ A schematic model for the interaction of Hsp40 and Hsp70 with a non-native protein. The substrate interacts with the dimer of Hsp40. Hsp40 transfers the substrate to Hsp70 and is released when the ATP is hydrolyzed. The nucleotide exchange by NEF leads to the release of the substrate.

binds, the hydrophobic linker region of Hsp70 binds to a groove of NBD, which opens the lid (SBD α) allowing a substrate to bind to SBD β . Hydrolysis of ATP leads to the release of the linker from NBD and to closing of the SBD by the lid (Figure 12.9).

12.1.2.4 Hsp90

Hsp90 is a dimeric chaperone that activates many regulatory and signaling proteins including oncogenic protein kinases and the tumor suppressor p53. It is an important target for the development of cancer drugs. It works at late stages of protein folding, sometimes together with Hsp70/Hsp40. Hsp90 has a large number of co-chaperones, among them the PPIase FKBP52.

The elongated subunit has three domains (Figure 12.10). The middle domain (MD) connects the N (NTD) and C (CTD) domains and interacts with client proteins. The ATP-binding site is at the NTD, which has the GHKL-fold (Gyrase, Hsp90, histidine Kinase, MutL). This fold has a four-stranded antiparallel β -structure, where ATP binds on top of the β -sheet (cf. Section 8.3.1). The CTD is responsible for dimer formation. The extreme C-terminus has the sequence MEEVD, which mediates the interactions with numerous co-chaperones with a number of tetratricopeptide repeats (TPR). Each such repeat is 34 residues long, forming a pair of α -helices.

The conformation without nucleotide can have an open V-shape, but this is an extreme, with more closed conformations coexisting. When ATP binds, the N domains interact tightly through a domain swap of the N-terminus and the subunits get into close contact. Upon ATP hydrolysis, the N domains lose their contact and the structure opens

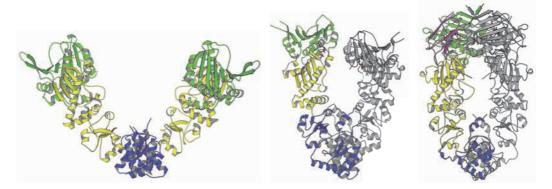


Fig. 12.10 ■ Three different conformations of Hsp90, depending on the nucleotide bound. The C-terminal domain (blue) at the bottom keeps the dimer interaction. The nucleotide-binding site can be seen at the lower part of the green domain. Left: The open enzyme from E. coli with no bound nucleotide (PDB: 2IOQ). Middle: The partly closed enzyme with ADP (PDB: 2O1V). Right: The closed yeast enzyme with ATP (PDB: 2CG9). The C-terminal domains (green) interact. The amphipathic loops are located in the hollow space between the M domains (yellow).

up partially (Figure 12.10). An amphipathic loop interacting with substrate or "client" proteins is part of the M domain in the cavity between the Hsp90 monomers.

The substrate or "client" protein can bind to the open nucleotide-free conformation. When ATP is bound, the N domains will dimerize to close the space between the M domains and assist in the conformational change of the substrate protein. Upon ATP hydrolysis, the N-terminal dimerization is lost and the dimer opens and releases the activated "client" protein (Figure 12.11, top).

Among many cochaperones, HOP binds at an early stage. It is a TPR protein and facilitates "client" transfer from Hsp70 (Figure 12.11, bottom). Another cochaperone is p23, which inhibits ATP hydrolysis and binds to the N domain when ATP is bound to trap the substrate protein in an active state. The release of p23 can be due to conformational changes of the substrate or interactions of other co-chaperones.

12.1.3 Chaperonins

One group of proteins called chaperonins are oligomeric folding chambers that sequester proteins with non-native folds. The most well known is GroEL (Hsp60) that belongs to group I chaperonins. Members of group I are primarily homooligomeric and occur mainly in bacteria, mitochondria and chloroplasts. Members of group II are normally heterooligomeric and found in archaeal or eukaryotic cytosol. The chaperonins consist of two rings, each with seven, eight or nine subunits. GroEL has seven-membered rings while the archaeal thermosome and the human TRiC has rings of eight, but rings of nine subunits also occur.

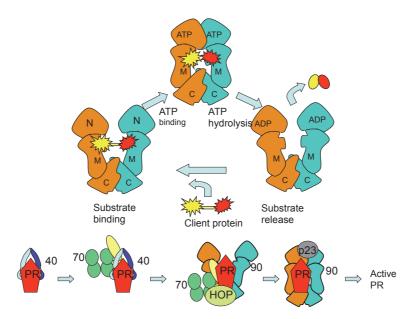


Fig. 12.11 ■ Top: A functional model for Hsp90. In apo-Hsp90, the dimer has an open form to which client proteins can bind. ATP induces a closed conformation where the N domains interact. Upon ATP hydrolysis, the interaction between the N domains is lost and the activated substrate protein is released. Bottom: Hsp90 can interact with Hsp40/Hsp70 in folding "client" proteins, in this case, the progesterone receptor (PR). The HOP cochaperone can bind both Hsp70 and Hsp90 in such a way that the substrate protein can be transferred from Hsp70 to Hsp90.

12.1.3.1 *GroE*

The GroEL chaperonin, also called Hsp60, is part of the GroE protein complex, which has a second component: the co-chaperonin GroES (Hsp10). GroEL is composed of two rings, each with seven subunits. Each ring forms a large chamber where a non-folded protein can bind. GroEL has three domains: the equatorial (E), the intermediate (I) and the apical (A) domains. The I domain functions like a hinge between the other two domains. The N- and C-termini are both situated in the ATP/ADP binding E domains forming the interface of the two rings. The A domains are at both ends of the GroEL cylinder (Figure 12.12).

GroEL assists the folding of a broad range of substrate proteins. The tip of the A domain is hydrophobic in nature and can bind and capture unfolded or partially folded substrate proteins at the entrance of the cylinder. Hydrophobic patches on two helices are the main points of contact. The large hydrophobic contact area may induce further unfolding of the bound substrate protein. The bound substrate induces binding of ATP. With ATP bound, the conformation of the apical domain changes to facilitate the binding of the GroES that acts as a lid of the cylinder (Figure 12.14). In this way, substrate proteins get trapped in the chamber, capped by GroES. As an effect of binding ATP and GroES, the

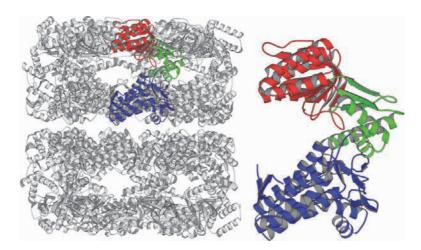


Fig. 12.12 ■ The structure of GroEL from E. coli without bound nucleotide. Left: The two sevenmembered rings are shown. One of 14 identical subunits is colored. Right: The detailed structure of the GroEL subunit with the equatorial domain (E, blue), the intermediate (I, green) and the apical (A, red) (PDB: 1XCK).

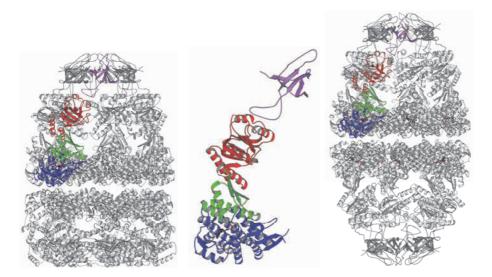


Fig. 12.13 ■ Left: The structure of GroEL from E. coli with ADP. One GroES (magenta) composed of seven subunits binds as a lid to the cis-ring and alters its structure. The A domain moves into an upright position to interact with GroES. The trans-ring is hardly affected. This structure is referred to as the "bullet model". Middle: The interaction between single subunits of GroES and GroEL (PDB: 1SX4). Right: The "football" structure of (GroEL-GroES)2 (PDB: 3WVL).

cavity becomes larger and the hydrophobic residues are hidden in their interaction with GroES. Instead, hydrophilic residues, primarily negatively charged ones, are exposed to the inner part of the cylinder (Figure 12.14). The substrate protein will then experience a drastic change of environment from hydrophobic to hydrophilic and in a wider chamber. The protein is then forced to refold on its own inside the chamber.

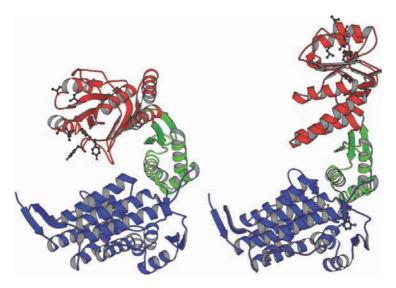


Fig. 12.14 ■ The two main conformations of the GroEL subunit. Hydrophobic side chains involved in substrate binding are shown as ball-and-stick models. Left: Non-liganded GroEL (PDB: 1XCK), Right: GroEL in complex with ADP and GroES (PDB: 1SX4).

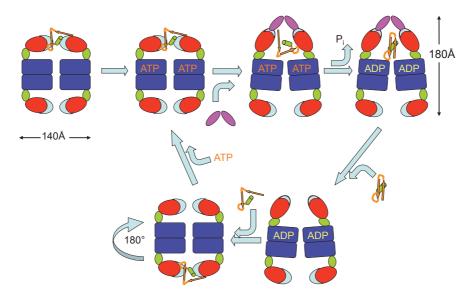


Fig. 12.15 ■ The functional cycle for GroE in the bullet model. The two rings have a negative cooperativity towards each other in binding ATP and non-native proteins. When GroES binds to the *cis*-ring, the opening to the *trans*-ring enlarges and can more easily bind new substrate proteins. The former *trans*-ring becomes the new *cis*-ring.

Two models of the functional cycle have emerged from in vitro experiments, but it is not clear if both operate in vivo. In the classical "bullet" model, GroES binds to alternate sides of the double barrel of GroEL, one at a time (Figure 12.15). In the more recent "football" model, GroES can bind to both sides of the GroEL barrel (Figure 12.16).

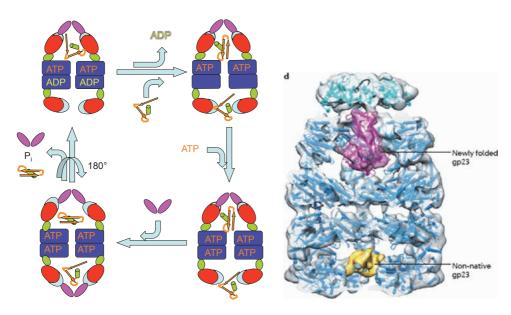


Fig. 12.16 ■ Left: The functional cycle of the football model. Note the 180° rotation of the pictures on the very left. Right: Cryo-EM structure of GroEL-GroES complex cut open to show the density of folded and non-native substrate proteins in the two rings. (Reproduced with permission from Saibil H. (2013) Chaperone machines for protein folding, unfolding and disaggregation. Nat Rev 14: 630–642. Copyright (2003) Macmillan Publishers limited.)

In the bullet model, the binding of ATP to one subunit stimulates the ATP binding to the other subunits in the same (cis)-ring (positive cooperativity), whereas it prevents ATP from binding to the opposite (trans)-ring (negative cooperativity). In this process, the subunits go through a range of conformational changes. GroES is then stimulated to bind to the cis-ring, but prevented from binding to the trans-ring (Figure 12.15). The binding of GroES leads to a large conformational change of the GroEL subunits. The apical domain becomes more upright to interact with GroES. The trans-cylinder of the two-cylinder engine is now ready to go through the same steps. When non-native polypeptides and ATP molecules bind to the trans-cylinder of GroES, the folded peptide is released from the former cis-ring. The former trans-ring now becomes the cis-ring.

The football model has emerged since symmetric complexes of (GroEL-GroES)₂ have been observed both by crystallography and cryo-EM (Figure 12.13). In addition, it may be possible that ATP association occurs before substrate binding. The football model is now considered as an intermediate in the functional cycle of GroEL-GroES. The dissociation of GroES from the complex seems to occur in a random manner. (Figure 12.16).

Regardless of the model, a protein that has not been properly refolded by one cycle of binding to GroE can repeatedly go through this process. GroE may also stimulate unfolding of misfolded proteins to allow them to refold to the native state.

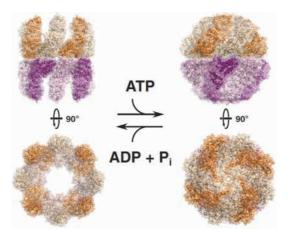


Fig. 12.17 ■ The structures of the archaeal group II chaperonin MmCpn in an open form without nucleotide (left) and closed with ATP bound (right) (PDB: 3IYF, 3LOS). Reproduced with permission from Lopez T, Dalton K, Frydman J. (2015) The mechanism and function of group II chaperonins. J Mol Biol 427: 2919–2930. Copyright (2015) Elsevier.)

12.1.3.2 Group II chaperonins

In archaea and eukaryotic cytosol, group II chaperonins help non-native proteins to fold properly. The relationship to GroEL-GroES is obvious, but the number of subunits in a ring can be eight or nine. Furthermore, in eukaryotes CCT (or TRiC) has eight subunits with distinct amino acid sequences (CCTα-1, CCTβ-2, CCTγ-3, CCTδ-4, CCT ϵ -5, CCT ζ -6, CCT η -7 and CCT θ -8). Group II chaperonins lack a cochaperonin like GroES, but they have an extension of the apical domain including the protrusion helix. The allosteric interactions in each ring of group II chaperones are in operation but allosteric interaction between the rings seem to be non-existent. Depending on the state of the bound nucleotide, the chaperone changes between an open substrate-binding conformation and a conformation which is closed, trapping the substrate in the folding chamber (Figure 12.17).

12.1.4 Folding During Protein Synthesis

When a newly synthesized protein emerges from the exit tunnel of the ribosome (Section 11.2.2.1), a number of protein factors compete to interact with the peptide. Among these are peptide deformylase, methionine amino-peptidase, signal recognition particle and the chaperone trigger factor (TF). The crowded environment in the cell requires a rapid and proper folding of the nascent chain to prevent degradation or unwanted aggregation. The exit tunnel of the ribosome is quite narrow and cannot allow any significant folding of the nascent polypeptide. Some proteins fold spontaneously

when they emerge from the exit tunnel. However, in many cases, chaperones are needed for the proper folding of the emerging polypeptide. Some chaperones interact directly with the ribosome and the nascent chain both in prokaryotic and eukaryotic systems. In bacteria, the nascent chain interacts primarily with small or "holding" chaperones that bind to ribosomes. TF and Hsp70 (DnaK) belong to this class of chaperones and are monomeric proteins that primarily prevent the aggregation of the growing polypeptide.

12.1.4.1 Trigger factor

Trigger factor is a bifunctional protein and belongs to the class of Hsp70-like chaperones but is also a peptidyl-prolyl-*cis/trans* isomerase (PPIase). It has a central role in identifying nascent peptides that need assistance to be folded and peptides that are addressed for export.

TF is composed of three domains: the N-terminal domain, the middle domain (the PPIase) and the C-terminal domain (Figure 12.18). The ribosome-binding site and the PPIase are at opposite ends of the molecule. The PPIase domain is dispensable and its role is not clear. TF is a dimer with interactions along the length of the elongated protein. However, it is a monomer when it binds to the ribosome. The N-terminal domain of TF binds close to ribosomal proteins uL23 and uL29 at the opening of the exit tunnel on the external surface of the large subunit.

TF binds to hydrophobic patches of substrate peptides in a groove along its entire length. The dimerization surface of TF coincides with its peptide-binding surface.

The total length of specific peptides in the peptide exit channel of the ribosome, with the added length of the trigger factor, suggests strongly that the peptide is almost

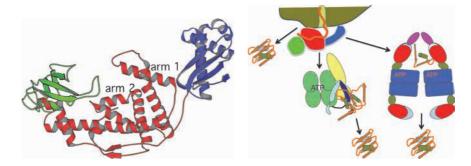


Fig. 12.18 ■ *Left*: The structure of the trigger factor (TF) from *Vibrio cholerae*. Trigger factor is a ribosome-bound chaperone. The N-terminal, ribosome-binding domain is to the right (blue) followed by the C-terminal domain (red) and the middle PPIase domain (green). The substrate binding is in a groove between the two arms (PDB: 1W26). *Right*: The folding of a nascent chain in bacteria emerging from the ribosome exit tunnel. The protein may be able to fold on its own or with TF. Alternatively, Hsp40 and Hsp70, as well as GroEL-GroES, may be needed to assist the folding of the nascent polypeptide.

completely extended. The fact that the nascent polypeptide is advancing step by step in TF during its synthesis suggests that the binding cannot be strong. In addition, the binding area is an open groove. TF probably forms a good protection against incorrect folding and degradation, but when enough amino acids are bound to TF to make a partial fold of the protein, they will detach from the groove and complete their folding.

12.1.4.2 Assisted folding of the nascent chain

In bacteria, the interactions with TF, Hsp40 and Hsp70 are normally sufficient for the proper folding of the protein (Figure 12.18). In other organisms, interactions with more complex chaperonins like GroEL-GroES are needed for the proper folding of the protein. Chaperonins are not known to interact with the ribosome. In eukaryotes, many chaperones may be needed to complete the folding process.

12.2 Folding, Unfolding and Degradation

As described in the previous sections, a range of proteins can assist in folding of polypeptides. This part of the chapter will extend on this theme but with a main focus on the degradation of proteins needed for protein homeostasis. Before a protein can be hydrolyzed to shorter peptides or amino acids, folded proteins or aggregates have to be dissolved to make them accessible to proteolytic activities. Therefore, there are many complexes that combine the two functions unfolding and proteolysis (Figure 12.19).

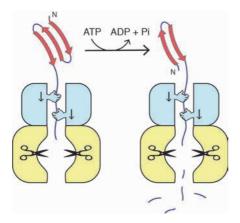


Fig. 12.19 ■ A schematic illustration of degradation of a protein by a combined machine for unfolding and proteolysis.

12.2.1 Protein Degradation

Protein turnover is a central feature of living organisms. The degradation of proteins is as important as their synthesis. A living cell needs a constant supply of amino acids for the synthesis of new proteins. These amino acids can come from breakdown of dietary proteins as well as from the degradation of cellular proteins. Thus, there are extracellular as well as intracellular proteolytic functions. Eukaryotic cells have organelles called lysosomes that degrade macromolecules, including proteins. The degradation of dietary proteins is mostly done in the gut by extracellular proteolytic enzymes of the digestive system such as pepsin, trypsin and chymotrypsin. In simple organisms, similar needs are met by extracellular proteases providing the supply of amino acids from degrading external proteins. These proteases are represented in brief (Table 12.2), since they are covered extensively in biochemical textbooks.

Cellular proteins have very different lifetimes and these are related to their function. Some proteins have long-term duties, while for others it may be just a few minutes. Different proteolytic activities are responsible for regulating the lifetime of a protein.

The intracellular proteases have two essential control roles: regulatory and quality. Proteolysis has to be specific and thoroughly controlled to prevent random destruction of cellular proteins. The proteolytic functions can be:

- Regulated by enzymes that need to be specifically activated, as in the blood coagulation system.
- Inhibited by various cellular protein inhibitors.
- Inaccessible to properly folded proteins.

The breakdown of cellular proteins is done by non-specific proteases, which are mostly oligomeric. These proteases are designed to control their activities so that only certain groups of substrates are degraded. Through the oligomeric structure, the active sites can be made inaccessible to folded proteins. When proteins expose a degradation tag (degron) proteolytic enzymes will identify them. These degrons are hidden in the correctly folded protein, but they get exposed when misfolded. Some degrons can be added as signals to

TABLE 12.2 Some Classes of Monomeric Proteases. For Further Information, See the MEROPS Database

| Type of Protease | Active Site Residues | Sequence Motifs | Example | |
|-------------------------|-----------------------------|------------------------|--------------------|--|
| Ser (trypsin type) | Ser, His, Asp | _ | Trypsin | |
| Ser (subtilisin type) | Asp, His, Ser | _ | Subtilisin | |
| Cys | Cys, His | _ | Papain | |
| Acid | Asp, Asp | _ | Pepsin | |
| Zn | Zn ligands: His, His, Glu | HEXXH | Carboxypeptidase A | |

the protein, like ubiquitin (Section 12.2.3). These labeled proteins are then directed to the protease or proteasome to be degraded to fragments that can subsequently be broken down to amino acids by other enzymes. During infections, the proteasome generates antigenic peptides that are presented by the MHC molecules on the surface of the cells (Chapter 17). These foreign peptides signal that the cell is infected and should be destroyed.

Non-natively folded proteins are a severe health problem, since they may aggregate (Section 3.3.3). Refolding by chaperones and degradation by proteolytic activities represent two main routes to handle misfolded proteins. There are functional and structural relationships between these routes. Many oligomeric intracellular proteases are ATPases but a large group is also ATP independent.

12.2.2 Oligomeric ATP-regulated Peptidases

Several large oligomeric peptidases need ATP for their activity. They are intracellular enzymes with their active sites enclosed in barrels, so they prefer to degrade unfolded peptides, leaving most proteins untouched. The ATPase activity involved in unfolding the protein to be degraded is performed by a common type of domain, an AAA+ module, as in a number of chaperones (Section 12.1.2.2). The AAA+ proteins are chemo-mechanical enzymes that can change the conformation of other proteins (Section 8.3). The central pore of the AAA+ domain is essential for the translocation of substrate proteins into the proteolytic chamber. A highly conserved aromatic residue, most often a tyrosine on a loop in the pore is involved in this translocation. The proteolytic activity can be performed by a domain of the protein or by separate subunits (Table 12.3 and Figure 12.20). These two

| Protein | Number of Protein Rings | Oligomeric Structure of AAA+ ATPase | Oligomeric Structure of Protease | Diameter of Entrance to Active Site | Active Site | | |
|-----------------------|-------------------------------|---|--|---|------------------|--|--|
| LonA/B | 2 | 6 domains | 6 domains | 18 Å | Ser+Lys | | |
| FtsH | 2 | 6 domains | 6 domains | 20 Å | Zn | | |
| HslVU | 4 | 6 HslU (or ClpY) sub- units on each side of the protease barrel | 6 × 2 HslV (or ClpQ) subunits | 20 Å | N-term Thr | | |
| Hsp100, ClpA, ClpB | 2 | 6 subunits | _ | 15 Å | _ | | |
| ClpAP/ClpXP | 4 | 6 ClpA/ClpX subunits on each side of the protease barrel | 2×7 ClpP subunits | 10 Å | Ser, His, Asp | | |
| 26S proteasome | 6 | 6 subunits | $7\alpha+7\beta+7\beta+7\alpha$ subunits | 13 Å | N-term Thr | | |

TABLE 12.3 Oligomeric ATP-Dependent Proteases

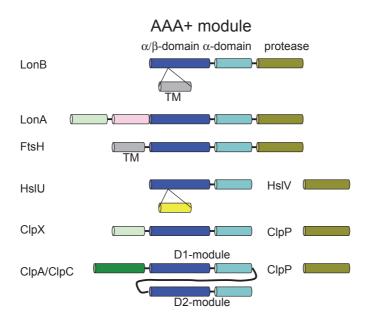


Fig. 12.20 ■ The domain organization of some of ATP-dependent proteases. Some are composed of one polypeptide and others of two, but all have an AAA+ module.

seemingly opposing activities to unfold the protein, in one case to be degraded, in the other case to be renatured, both consume ATP.

AAA+ modules are composed of two domains: the α domain (smaller) and the α/β domain (larger; Figure 12.20). The nucleotide binds to elements of both domains. Helix 1 of the α domain is a crucial element. It can be straight, bent to variable extents, or contain a two-residue bulge in the middle. This affects the space and the angle between the α domain and the α/β domain. Furthermore, the relationship between the α domain and the α/β domain depends on the state of the adenine nucleotide, which binds between subunits. While the nucleotide binding is to both domains of one subunit, the Arg finger (Section 8.3) comes from a neighbor subunit. ATP hydrolysis leads to conformational changes of the hexameric ring, which probably leads to pulling the substrate peptide or protein into the degradation chamber. The hexameric nature of the AAA+ components could suggest that the subunits undergo ATP hydrolysis and ADP -> ATP exchange sequentially, as does ATP synthase (Section 5.3).

The unfolding ATPase ClpX, like its close relative ClpA, is involved with the protease ClpP. Its subunits do not operate in a symmetric fashion. Only four of the identical subunits in ClpX are loadable (L) with ATP, while two of them are not. In the unloadable (U) subunits, the small and large domains have a conformation that prevents nucleotide binding. The organization of the subunits in the ring is L-U-L-L-U-L (Figure 12.21). Nevertheless, the small domain of one subunit and the large domain of the next subunit always interact in the same way, a static interaction.

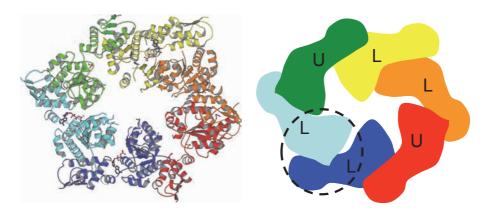


Fig. 12.21 ■ The asymmetric arrangement of loadable (L) and unloadable (U) AAA+ subunits in ClpX. The rigid interaction between the small domain of one subunit and the large domain of the next domain is also shown. The black line identifies the rigid contact between a loadable and unloadable subunit (PDB: 4I81).

12.2.2.1 Hsp100

The Hsp100 (eukaryotes) or ClpB (bacteria) proteins are involved both in protein disaggregation on their own, or degradation when combined with proteolytic proteins. The Hsp100 family includes Hsp104, ClpA, ClpB, Hsp78 and are homo-hexameric oligomers with two AAA+ domains each (Figure 12.22 and Section 8.3). The role of the double rings remains unclear.

The N-terminal domain (NTD), which is loosely hinged to the main body of the hexamer (Figure 12.22), is involved in substrate recognition mediated by small adaptor proteins like ClpS. ClpS directs the activity of Hsp100 towards aggregated proteins. NTD is composed of two helical bundles with four helices each, related by a pseudo-two-fold axis, and has a hydrophobic surface. ClpS interacts with NTD through several contacts. The sticky surface of NTD probably contributes to unspecific binding of substrates, while the functional binding occurs on the AAA - 1 or AAA - 2 domains (Figure 12.22). The width of the central channel through which the proteins are pulled to regain structure or getting ready for degradation is around 15 Å. The conserved tyrosine residue in the central channel of the AAA – 1 or D1 ring is important for the disaggregation activity of ClpB. If tyrosine is mutated and the NTD is removed, then activity is lost.

12.2.2.2 *HslVU*

In some bacteria and eukaryotes, there is a distant relative of the proteasome (see below). The name of this group of proteases is HslV (heat shock locus V) or ClpQ. HslV is induced by heat and degrades proteins with a non-native fold. Similar to the 20S proteasome the particle is built of two hexameric rings forming a cavity for the proteolytic degradation (Figure 12.24). HslV has 20% sequence similarity to the β subunit of the proteasome, but

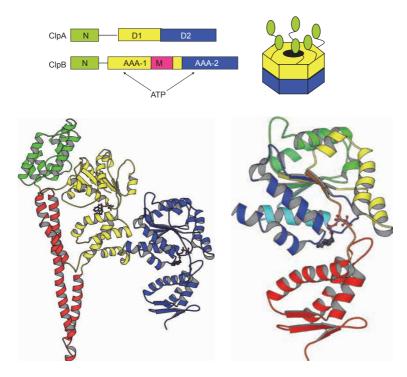


Fig. 12.22 ■ Top: Hsp100/ClpA and Hsp104/ClpB have double ATPase rings. Upper middle: The N-terminal domains of the hexamer (green) are bound through flexible linkers and the AAA+ domains form a six-fold double disk structure. Lower left: The structure of a monomer of ClpB from T. thermophilus. The red middle (M) domain, not present in Hsp100, forms wing-like structures on the surface of the hexamer. The colors correspond to the ones in the upper left picture. The two ADPNP molecules are shown as ball-and-stick models (PDB: 1QVR). Lower right: The C-terminal AAA+ domain of ClpB colored from blue (N-terminal) to red (C-terminal). The P-loop, which binds the nucleotide, follows the first strand of the sheet. The C-terminal helix bundle (red) is part of all AAA+ ATPase domains.

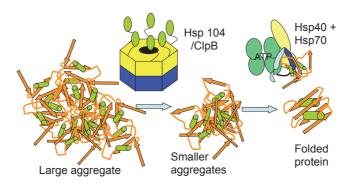


Fig 12.23 ■ A possible path in the disassembly of aggregates. Hsp104/ClpB breaks down larger aggregates to smaller ones that can be substrates for the Hsp70/DnaK family of chaperones that can generate a protein with native fold.

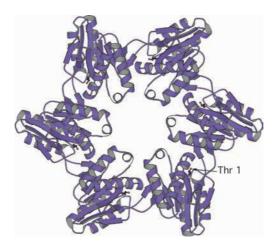


Fig. 12.24 ■ The structure of *Thermotoga maritima* HslV, a bacterial relative of the 20S proteasome. It is composed of two hexameric rings of identical subunits (only one of the rings is shown). The structure of the subunit is related to the β subunit of the proteasome and is a threonine protease (PDB: 1M4Y). The catalytic N-terminal residue (Thr 1) is marked.

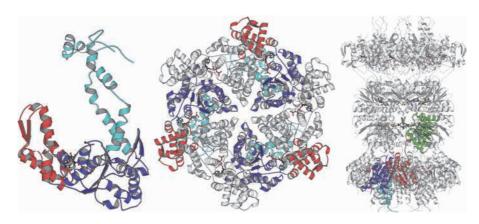


Fig. 12.25 ■ Left: The structure of ClpY (PDB: 1E94) with its three domains (N blue, I cyan and C red), seen in the monomeric form and as a hexamer (middle) with bound ATP. Right: The complex of HsIV and HsIU from E. coli. The two proteolytic HsIV rings (one green monomer) can be flanked by one or two ATPase HslU rings. The complete complex is a relative of the complete proteasome (PDB: 1KYI).

the proteasomal α subunits are lacking. HslV is an Ntn-hydrolase (N-terminal nucleophile hydrolases) with Thr1 as the nucleophile. The active site in the inner cavity has a low protease activity on its own on short peptides and negligible activity on proteins.

Like the 20S proteasome, HslV can associate with an ATPase of the AAA+ superfamily, called HslU or ClpY. HslU enhances the proteolytic activity of HslV by one to two orders of magnitude. HslU also forms hexameric rings and binds to each side of the double HslV rings (Figure 12.25). HslU is a member of the Clp/Hsp100 family of chaperones, but together with HslV is also engaged in the degradation of target proteins.

The HslU monomer has three domains: the N-terminal or the ATPase, an intermediate domain (I) and a C-terminal domain. The ATPase domains face away from the HslV subunits. The interaction between HslU and HslV induces conformational changes that may explain the increased activity of the complex. A complex of two rings of HslV with one ring of HslU shows an asymmetric rearrangement of the structure. The annulus through which the protein is translocated becomes much wider, almost 20 Å in diameter, upon binding of the HslU ring. The interface between the two HslV rings is hydrophobic, while the HslV-HslU interface is significantly hydrophilic with many hydrogen bonds. The N and C domains of HslU may be responsible for the assembly and activation of the complex, while the I domains may be involved in the selection and channeling of substrates into the proteolytic cavity of HslV.

12.2.2.3 FtsH

FtsH differs from other AAA+ proteases in several ways. It is bound on the inside of the cytoplasmic membrane, essential and conserved in bacteria, important in mitochondria and chloroplasts. Its role is quality control of membrane proteins. The N-terminal region has two transmembrane helices and a region is located in the periplasm (Figure 12.26). The transmembrane region, the AAA+ domain and the protease domain are all contained within the same polypeptide. The energy of ATP is used to pull non-native proteins out of the membrane to be unfolded and degraded.

The structure of FtsH is arranged as two rings (Figure 12.26). The protease ring is a flat hexagon with six-fold symmetry, composed of a helical domain. The active sites are enclosed between the rings and have a zinc ion bound by two histidine and one aspartate

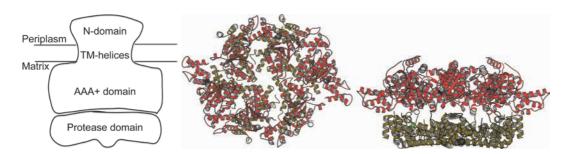


Fig. 12.26 ■ *Left*: The full structure of FtsH as a schematic drawing. *Middle*: Looking down onto the AAA+ domain. *Right*: The AAA+ domain and the protease domain seen from the side. The two rings are built of six polypeptides. The AAA+ domain (red) is on top and the protease domain (brown) is at the bottom. The active proteolytic sites are in the inner chamber. The AAA+ domain is connected to the cell membrane by two *trans*-membrane helices per subunit (PDB: 2CE7).

residues. The histidines are part of a frequently occurring zinc-binding motif of proteases or "zincins", HEXXH. The active sites in the hexamer are about 34 Å apart.

The AAA+ domains form a toroid. While the protease ring is an almost perfect hexamer, the six AAA+ domains have two-fold symmetry. The width of the pore between the AAA+ domains is 20 Å. The AAA+ ring is likely to undergo conformational changes and changes of symmetry during binding and hydrolyzing of ATP and pulling in and denaturing target proteins. The aromatic residue involved in the pulling activity is a phenylalanine seen at three different levels at the central pore of FtsH in the static crystal structure.

12.2.3 The Ubiquitin Pathway and Proteasomes

Two major proteolytic systems of the eukaryotic cell are the lysosome and the proteasome. The lysosome is a cellular organelle with a large set of enzymes able to degrade all kinds of biomolecules. Proteasomes are large oligomeric aggregates responsible for intracellular breakdown of proteins into smaller fragments that will subsequently be degraded to amino acids by smaller proteases. The proteasome handles proteins that are marked for destruction by their covalent linkage to ubiquitin. This system has been called the "garbage disposal". Malfunction of the ubiquitin-proteasome system can lead to accumulation of misfolded and aggregated proteins, which frequently leads to neurodegenerative diseases.

12.2.3.1 The ubiquitin labeling system

Ubiquitin (Ub) is a small and stable protein of 76 amino acids. It can be covalently attached to proteins as a signal for various cellular purposes. Ub has a Gly-Gly sequence at the flexible C-terminus that is the attachment point to the target protein. Three different proteins (E1, E2 and E3) are engaged in the labeling process (Figure 12.27). A ubiquitin-activating protein (E1) adenylates the C-terminus of Ub and subsequently transfers it to a cysteine of E1 through a thioester linkage. Subsequently, E1 transfers the Ub moiety to a cysteine of an ubiquitin-conjugating enzyme (E2). Thus, E1 carries out three distinct reactions: activation of ubiquitin by adenylation, thioesterification and transthioesterification. The enzyme is quite complex, with two Rossmann-fold domains (one active and one inactive), a cysteine domain and an ubiquitin-fold domain. Ubiquitin binds to the active Rossmann-fold domain. On the other hand, E2 are relatively small single domain proteins. Ubiquitin-protein ligases (E3) finally transfer the ubiquitin to a substrate protein that will be covalently labeled, through an isopeptide bond between one of its lysyl residues and the C-terminal Gly of Ub. The E3 molecules are specific for different types of

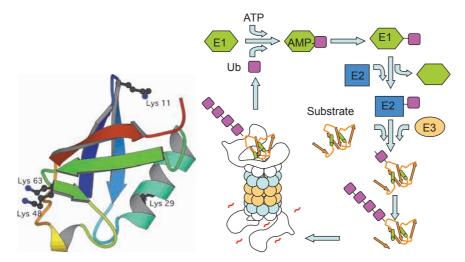


Fig. 12.27 ■ Left: The ubiquitin molecule (Ub; PDB: 1UBQ). Right: The steps involved in the labeling of a substrate protein (red) with Ub (purple). E1 (green) is the ubiquitin-activating protein. It first activates the Ub by transferring it to the AMP part of an ATP molecule and subsequently to a cysteine residue of E1. The ubiquitin-conjugating enzyme (E2, blue) then becomes labeled with the Ub. Finally, the ubiquitin ligase (E3, light brown) transfers the Ub to the amino group of a lysine of the substrate protein. In some instances, the E3 enzyme first covalently binds the Ub. Substrates that are destined for degradation become polyubiquitinated, bind to the proteasome and unfold. This leads to the translocation of the substrate protein into the proteolytic chamber and its degradation. The Ub moieties are released and reutilized.

target proteins and can be grouped into three different classes called RING, HECT and RBR. The E2 conjugating molecules are specific for certain groups of E3. While there are only two E1 proteins in a mammalian cell, there are around 40 E2 variants and over 600 E3 enzymes.

Ub labeling has numerous different functions in the cell, one being a signal for degradation by the proteasome. Ub can be attached to one or several lysine residues of a protein and whole chains of Ub molecules can be attached to a specific lysine. This polyubiquination can be linear or branched. When a substrate protein is polyubiquinated, this is done through initiation and elongation steps. As a rule, four Ub molecules are needed as a signal for degradation by proteasomes (Figure 12.27). However, for small substrate proteins or peptides, even one Ub is enough as a signal for degradation. There are also numerous ubiquitin-like proteins that are ligated to various protein substrates, but they normally have other cellular roles than the labeling for destruction. SUMO (small ubiquitin-related modifier) is probably the ubiquitin-like protein that has received the greatest attention since it labels different transcription factors.

Ub has seven lysyl residues, four of which (K11, K29, K48, K63) are involved in isopeptide linkages to the C-terminal glycine of the next ubiquitin in polyubiquitin. The linkage through K48 and K29 of polyubiquitin is the signal for degradation by the proteasome. Monoubiquitination of target proteins has different cellular roles.

12.2.3.2 Proteasomes

Proteasomes are found in all branches of life. They are essential for all eukaryotic cells but not essential for the viability of archaea, and are found only in some eubacteria. Proteasomes are the sites for regulated degradation of intracellular ubiquitinated proteins to oligopeptides. The substrates can be short-lived regulatory proteins or non-natively folded proteins. Normally, the peptides are 7–9 residues long after degradation by the proteasome. These peptides can be further processed by other proteases.

The full proteasome is a 26S particle. The molecular mass is 2.5 MDa. The eukaryotic 26S proteasome is composed of at least 34 different subunits. One or two 19S regulatory particles extend from the central structure, the core, or the 20S proteasome. The 20S particle is a cylinder with a height of about 150 Å and a diameter of 110 Å, composed of four heptameric rings. The outer two rings are built up of α subunits while the inner two rings are oriented head-to-head and built up of 14 β subunits.

The α and β subunits have a common fold, with a sandwich of two five-stranded β -sheets with two α -helices on each side (Figure 12.29). The H1 and H2 helices are responsible for the contact between α and β subunits, whereas the H3 and H4 helices form the contacts between the β -rings. The ring of β subunits form a chamber, and the α subunits form a narrow annulus at the entrance of the chamber. The proteolytic active sites are found on the inner side of the β subunits (Figure 12.28). Thus, there can be 14 active proteolytic sites in a proteasome.

All bacterial and most archaeal proteasomes are built up of only one type of α and one type of β subunits. For some archaea and eubacteria, there are two different α and two different β subunits. In eukaryotes, there are seven different but homologous subunits each of the α and the β type (Figure 12.30). Thus, in eukaryotic 20S proteasomes, the exact seven-fold symmetry is broken.

The complete proteasome can degrade both native folded proteins that are ubiquitinated as well as non-natively folded proteins. However, the 20S part of the proteasome can degrade only unfolded proteins. The annulus generated by the α subunits has a diameter of 13 Å, which is too narrow to allow a folded protein to enter. Thus, before the substrate can be translocated into the proteolytic chamber it needs to be unfolded. This activity is performed by the 19S complex and corresponds to chaperone activity on one or both sides of the cylinder. This regulatory complex prevents uncontrolled proteolysis and is responsible for recognition, deubiquitination, unfolding and translocation of the substrate into the chamber for degradation. It is composed of 20 subunits and can be dissociated into two units called the base and the lid (Figure 12.30). The base contains six ATPase subunits (Rpt1-6), which pull the substrate protein into the proteolytic chamber and belong to the superfamily of AAA+ enzymes. Each subunit has a unique amino acid

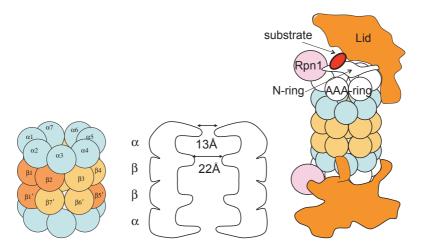


Fig. 12.28 Schematic drawings of the 26S proteasome. *Left*: The central 20S core (colored) is composed of four rings, each with seven subunits. The two β-rings (brown) are in the center and the α-subunits (blue) form the outer rings. In the eukaryotic 20S proteasomes, the seven α and the seven β subunits are all unique. The figure illustrates the spatial relationship between the active β subunits (darker brown) 1, 2 and 1′ on one side of the proteolytic chamber and 5′ and 5 on the opposite side. *Middle*: A narrow channel leads into the antechambers and subsequently into the central proteolytic chamber. The active sites face into the inner chamber formed by the β subunits. *Right*: In the 26S proteasome the 20S core is capped on both sides by the regulatory 19S complex composed of 20 subunits. The base part is closest to the 20S core and the lid forms the outer portion.

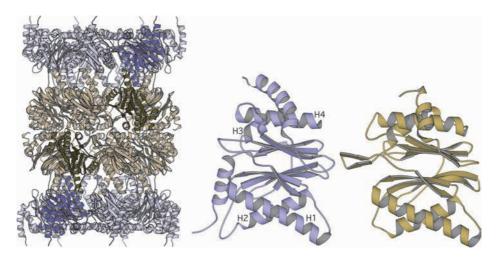


Fig. 12.29 ■ *Left*: The crystal structure of the 20S proteasome from *Archaeoglobus fulgidus. Right*: The α and β subunits have very similar structure (PDB: 1J2Q).

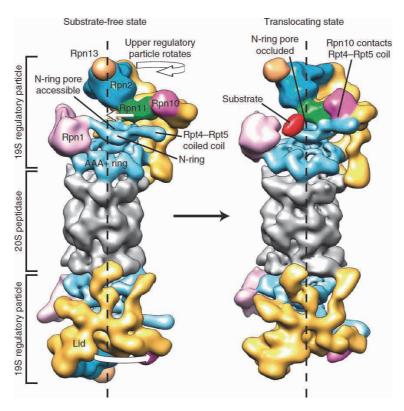


Fig. 12.30 ■ A cryo-EM structure of the complete 26S proteasome. The central 20S component is shown in gray. The 19S regulatory complexes are attached to the 20S. The base (blue) forms two rings, the AAA+ ring and the N-ring. In the right structure, a substrate protein is bound. (Reproduced with permission from Matyskiela ME, Lander GC, Martin A. (2013) Conformational switching of the 26S proteasome enables substrate degradation. Nat Struct Mol Biol 20: 781-788. Copyright Nature America Inc (2013).)

sequence and a fixed position in the ring. The N-terminal small domains of the AAA+ have the oligomer-binding fold (OB) and form a separate ring above the larger domain forming the ATPase ring. C-terminal tails of these six subunits interact specifically with the seven α subunits. The base also contains a ubiquitin receptor Rpn13. In addition, the base contains two large scaffolding proteins, Rpn1 and 2, and a deubiquinating enzyme, Ubp6.

The lid portion of the 19S complex is composed of subunits Rpn3, 5 – 7, 9, 10 and 12 forming a horseshoe-like structure. Its most important function, to remove the ubiquitin units from the substrate, is done by Rpn11, which forms a heterodimer with Rpn8. Rpn13 and Rpn10 are jointly involved in the recognition of the polyubiquitin chain attached to substrates.

Without a substrate protein the ATPase subunits are arranged as a fixed spiral staircase. Furthermore, the symmetry axis of the N-ring and the ATPase-ring in the apoproteasome does not superimpose with the symmetry axis of the 20S complex. However, with an ubiquinated substrate protein (Figure 12.31), a conformational change makes the pores superimpose and the ATPase-ring become essentially flat. However, the pore

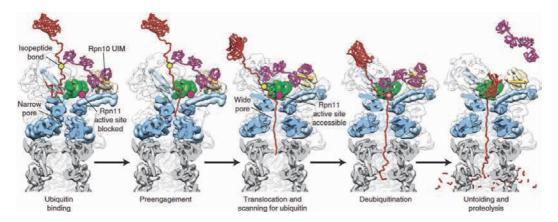


Fig. 12.31 ■ The binding and processing of an ubiquinated protein by the proteasome. (Reproduced with permission from Matyskiela ME, et al. (2013) Conformational switching of the 26S proteasome enables substrate degradation. Nat Struct Mol Biol 20: 781-788. Copyright (2013) Nature America Inc.)

loops involved in the translocation of substrate remain in a spiral staircase arrangement due to different tilting of the large domain. It is not known how the ATPase subunits and pore loops behave during the active burst of ATP hydrolysis and release of ADP and P_i. In binding a substrate, Rpn11 also shifts position from being off-axis to become on-axis. As in the case of ClpX, the interaction between the small domain of the AAA+ subunit and the large domain of the next AAA+ subunit functions as a rigid body (Section 12.2.2).

In mammals, there is an alternative type of proteasome called the immuno-proteasome. It is suggested to produce antigenic peptides to be presented by MHC class I molecules. This specialized proteasome has an 11S/PA26 complex bound to the 20S core (Figure 12.32). In the 20S, the N-termini of the α subunits block access to the chamber through the annulus. The seven-fold symmetry of PA26 orders the N-termini of the seven α subunits. This in turn has the effect that the annulus opens up to simplify the access to the proteolytic chamber. Another related system is the Blm10/PA200. Here, a large single chain activator wraps around the end of the 20S proteasome. The N-termini of the seven α subunits only get partly ordered, making access to the proteolytic chamber limited.

The 20S proteasome has three inner cavities with a diameter of around 50 Å (Figure 12.28). The central chamber formed by the face-to-face oriented $\boldsymbol{\beta}$ subunits contains the active sites. The entries to this inner chamber are around 30 Å, while the entry into the outer chambers is narrower. The interactions with the 19S regulatory complex open these outer annuli.

The proteasome β subunits belong to the class of Ntn-hydrolases. They are always produced as inactive precursors and converted to their active form by internal autocatalytic cleavage. This makes Thr1 (the N-terminal residue) able to act as the nucleophile, and the free amino group at the N-terminus acts as the general base for the hydrolysis.

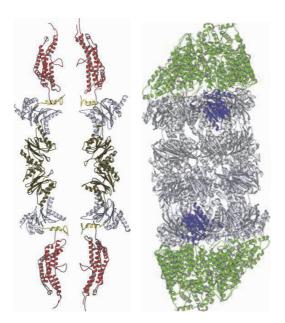


Fig. 12.32 ■ Left: A central section of the yeast 20S proteasome in complex with PA26, an 11S activator (red). The yellow parts are the ordered N-termini of the α subunits of the proteasome (PDB: 1Z7Q). Right: The crystal structure of the single chain of yeast Blm10 (green) in complex with the 20S particle (PDB: 4V7O). One of the 14 α subunits is shown (blue) in both double layers. The active sites are as always located in the interior of the central rings.

This free N-terminal amino group is the common feature of all Ntn-hydrolases. The threonyl hydroxyl oxygen becomes covalently bonded to the carbonyl carbon of the S1 residue of the substrate. In a following step, the bond is hydrolyzed.

The pro-peptides are of different lengths and sequences, but the residue preceding the threonine is always a glycine. For autocatalytic cleavage, there is no proton acceptor available. Therefore a water molecule is required. A number of conserved residues in the vicinity of the threonine participate in the interactions within the enzyme as well as with the substrate. The distance between active sites is around 28 Å, which may relate to the size of the peptides produced.

In eukaryotes, there are seven different β subunits, but only β 1, β 2 and β 5 are active enzymes. Thus, compared to bacteria where 2×7 subunits all are active, eukaryotic organisms only have 2 × 3 active subunits. These subunits bind aldehyde inhibitors. Mutation of Thr1 in one of these subunits results in the loss of at least one of the classical proteasomal activities. β1 preferentially cleaves after Glu, β2 cleaves after Arg or Lys (like trypsin) and β5 cleaves after aromatic residues like chymotrypsin. However, the specificity is not strict and the eukaryotic proteasome can cleave most peptide bonds. The active sites as well as the inactive subunits of both β-rings show a complex cooperativity in binding the substrates and guiding the possibilities for cleavage.

Proteasomes perform regulated degradation of key proteins in different pathways. Therefore, inhibitors of proteasomes are of significant interest. The cell cycle can be arrested at various stages, which leads to decreases in cell proliferation. Structural and functional analyzes of many inhibitors have been performed.

12.3 Serpins: Protease Inhibitors

There is a large set of proteins that function as protease inhibitors, which are very much needed to prevent proteolytic catastrophes. Among these inhibitors there are more than 20 families of serine protease inhibitors, of which the most abundant group in higher organisms are the serpins. This type of inhibitor is an example of an unusual structural activity, inserting a β -strand in the middle of an existing β -sheet.

The serpins are a large family of serine protease inhibitors that are found in all kingdoms of life. They are suicide inhibitors and consumed in a stoichiometric fashion together with the specific targets of their inhibition. Some members of this family are antitrypsin, antichymotrypsin (ACT), antithrombin, plasminogen activator inhibitor-1 (PAI-1), α 1-proteinase inhibitor (α 1-PI) and ovalbumin. Ovalbumin constitutes about two-thirds of the protein in egg white; it has the same general structure as the serpins, but does not function as an inhibitor. The serpins are all involved in tightly regulated proteolytic pathways in humans: blood coagulation, fibrinolysis, complement cascade, tissue remodeling, tumor metastasis, inflammation and apoptosis. Despite low sequence similarity, the serpin structures are highly similar to each other, with nine α -helices (A-I) and three β -sheets (A-C).

The serpin class of proteins fold to a metastable state that can become stable through proteolysis. They were first thought to be unique in this respect. However, now there are many other metastable proteins known. They fold into a metastable state that can be converted to a stable form by covalent or non-covalent interactions. Another example of metastable proteins is the prion proteins, which have a normal, metastable form that can be converted to the disease-causing stable form.

Native serpins are kinetically trapped in a high-energy state that can be relaxed in two different ways as shown below. The melting temperature for the metastable fold (the stressed state) is for a typical serpin 60° C, compared to 120° C for the stable form (the relaxed state). In the metastable form, the reactive centre loop (RCL) of between 20–25 amino acids is exposed and quite flexible. This loop can be incorporated in the structure in an unusual way; it becomes inserted as a new β -strand in the middle of β -sheet A (Figure 12.33). This type of structural rearrangement was not anticipated, but is now well characterized.

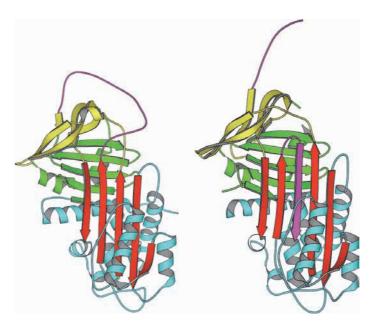


Fig. 12.33 ■ The structure of the metastable or stressed state (left, PDB: 1HP7) and the stable, relaxed state (*right*, PDB: 1QMB) of antitrypsin. The protein has three β-sheets, A (red), B (green) and C (yellow). The reactive center loop (RCL, purple) in the metastable form becomes cleaved and integrated in the middle of the A-sheet to become the stable relaxed state.

The RCL loop has a central function for serine protease inhibitor activity. Different proteolytic enzymes recognize the specific amino acid sequence of the loop as a substrate (Figure 12.34). They bind and catalyze the cleavage of the loop. However, the active site serine of the enzyme remains covalently linked to the N-terminal side of the cleavage site of the serpin. This leads to a large conformational change of the serpin, whereby the loop becomes inserted into β -sheet A. This means that the protease is translocated about 75 Å from the proximal to the distal side of the serpin. Crystal structures of complexes in this state show that parts of the protease have much increased flexibility, as indicated by crystallographic B-factors. NMR analysis of the trypsinantitrypsin complex suggests that the protease gets into a molten globule state. The flexible regions of the protease become susceptible to proteolysis, which leads to its destruction. However, the RCL of the serpin is already cleaved and it cannot be reused and the serpin is therefore also degraded.

Serpins can also undergo this strand insertion without cleavage of the RCL. This is a modulating mechanism by which the inhibitor is converted to a latent, inactive form. Figure 12.35 shows structures of a number of intermediate steps in inserting the RCL loop into the center of the β -sheet A, without cleavage of the peptide.

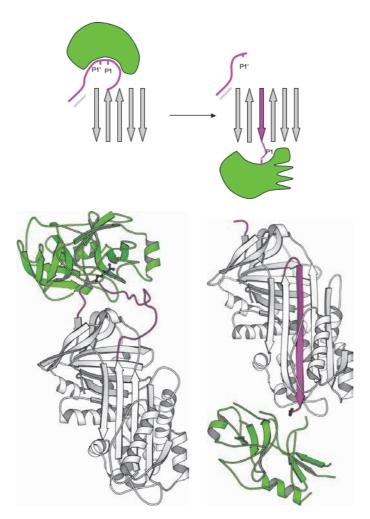


Fig. 12.34 Simplified (*top*) and schematic (*bottom*) drawings of the action of serpins. *Left*: The protease (green) binds to the (PDB 1OPH) loop of the serpin (purple), which is connected to one of the central strands in the sheet. *Right*: This leads to the hydrolysis and incorporation of the RCL loop as a β-strand in sheet A of the sorption (1EZX). As part of the process, the protease is swung to the other side of the serpin. The interactions cause a flexibility of the protease, which in turn makes it susceptible to protease degradation and elimination.

The insertion of RCL into the A-sheet of another serpin molecule is only one of the possible alternatives.

The metastable state obviously has structural elements that are not compatible with a highly stable protein. These elements include exposure of hydrophobic groups and burying hydrophilic groups without proper hydrogen bond matching. At the same time, the RCL contains elements that will fit excellently into the place of the middle strand in β -sheet A. Even a partial insertion of RCL into the sheet is an energetically favorable step.

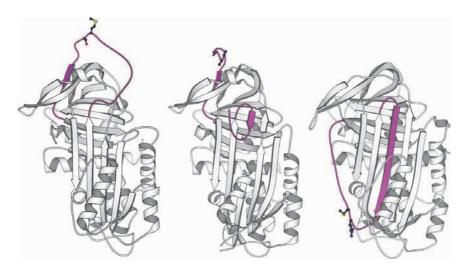


Fig. 12.35 ■ Structures illustrating the gradual and spontaneous incorporation of the RCL loop (purple) into the middle of a β-sheet A. The peptide bond in the loop is not cleaved, but the inhibitor is transformed to a latent inactive form. The side chains of the amino acids on both sides of the bond that would be cleaved in the active inhibitor are shown as a ball-and-stick model. From *left*: α_1 -antitrypsin/PDB: 1QLP/, δ form of α_1 -antichymotrypsin/PDB: 1QMN/ and plasminogen activation inhibitor PAI-1/PDB: 1LJ5/.

For Further Reading Section 12.1

Original Articles

Baram D, Pyetan E, Sittner A, et al. (2005) Structure of trigger factor binding domain in biologically homologous complex with eubacterial ribosome reveals its chaperone action. Proc Natl Acad Sci USA 102: 12017-12022.

Fei X, Ye X, LaRonde NA, Lorimer GH. (2014) Formation and structures of GroEL-GroES2 complex chaperonin footballs, the protein folding functional form. Proc Natl Acad Sci USA 111: 12775-12780.

Howard BR, Vajdos FF, Li S, et al. (2003) Structural insight into the catalytic mechanism of cyclophilin A. Nat Struct Biol 10: 475–481.

Lakshmipathy SK, Tomic S, Kaiser CM, et al. (2007) Identification of nascent chain interaction sites on trigger factor. *J Biol Chem* **282**: 12186–12193.

Lopez T, Dalton K, Frydman J. (2013) The mechanism and function of group II chaperonins. J Mol Biol 427: 2919–2930.

Xu Z, Horwich AL, Sigler PB. (1997) The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature 388: 741-750.

Reviews

- Haslbeck M, Vierling E. (2015) A first line of stress defense: Small heat shock proteins and their function in protein homeostasis. J Mol Biol 427: 1537–1548.
- Harrison CJ. (1997) La cage aux fold: Asymmetry in the crystal structure of GroEL-GroES-(ADP)7. Structure 5: 1261-1264.
- Krukenberg KA, Street TO, Lavery LA, Agard DA. (2011) Conformational dynamics of the molecular chaperone Hsp90. Quart Rev Biophys 44: 229-255.
- Mayer MP, Kityk R. (2015) Insights into the molecular mechanism of allostery of Hsp70s. Front Mol Biosci 2(58): 1–7.
- Saibil HR, Fenton WA, Clare DK, Horwich AL. (2013) Structure and allostery of the chaperonin GroEL. J Mol Biol 425: 1476–1487.
- Schiene-Fischer C. (2015) Multidomain peptidyl-prolyl cis/trans isomerases. BBA 1850: 2005–2016.
- Skjærven L, Cuellar J, Martinez A, Valpuesta JM. (2015) Dynamics, flexibility, and allostery in molecular chaperonins. FEBS Lett 589: 2522–2532.
- Wang I, Wang X, Wang C. (2015) Protein disulfide isomerase, a folding catalyst and a redoxregulated chaperone. Free Radic Biol Med 83: 305-313.
- Ünal CM, Steinert M. (2014) Microbial peptidyl-prolyl cis/trans isomerases (PPIases): Virulence factors and potential alternative drug targets. Microbiol Mol Biol Rev 78: 544-571.

For Further Reading (Section 12.2)

Original Articles

- Matyskiela ME, Lander GC, Martin A. (2013) Conformational switching of the 26S proteasome enables substrate degradation. Nat Struct Mol Biol 20: 781–788.
- Vostrukhina M, Popov A, Brunstein E, et al. (2015) The structure of Aquifex aeolicus FtsH in the ADP-bound state reveals a C2-symmetric hexamer. Acta Cryst D71: 1307–1318.

Reviews

- Ciechanover A. (2004) Intracellular protein degradation: From a vague idea, through the lysosome and the ubiquitin-proteasome system, and onto human diseases and drug targeting. Les Prix Nobel (ed. Frängsmyr T), 1197–1211.
- Ciechanover A, Stanhill A. (2014) The complexity of recognition of ubiquinated substrates by the 26S proteasome. BBA **1843**: 86–96.
- Huntington JA. (2006) Shape-shifting serpins Advantages of a mobile mechanism. TIBS 31: 427-435.
- Komander D, Rape M. (2012) The ubiquitin code. Ann Rev Biochem 81: 203–229.

- Lorenz S, Cantor AJ, Rape M, Kuriyan J. (2013) Macromolecular juggling by ubiquitylation enzymes. BMC Biol 11: 65.
- Nyquist K, Martin A. (2014) Marching to the beat of the ring: Polypeptide translocation by AAA+ proteases. TIBS 39: 53-60.
- Saibil H. (2013) Chaperon machines for protein folding, unfolding and disaggregation. Nat Rev 14: 630-642.
- Sauer RT, Baker TA. (2011) AAA+ proteases: ATP-fueled machines of protein destruction. Ann Rev Biochem 80: 587-612.
- Whisstock JC, Bottomley SP. (2006) Molecular gymnastics: Serpin structure, folding and misfolding. Curr Opin Struct Biol 16: 761-768.

Database for Proteases

http://merops.sanger.ac.uk



Transmembrane Transport

13.1 Types of Protein-Catalyzed Transmembrane Transport

A fundamental aspect of biomembranes is their ability to control transport and signal transduction through the activity of membrane transport proteins. Cellular transport systems are usually highly specific and are subject to various modes of regulation, and different kinds of membrane proteins facilitate these different types of transport. The direction of solute transport can be *against* electrochemical gradients and therefore *active* (i.e. energy-consuming). This kind of transport is facilitated by primary and secondary transporter proteins. Or, transport can be *down* electrochemical gradients and therefore *passive* (i.e. does not require the input of energy). This is the case for pores and channels.

There are three major families of membrane transport proteins, and these are illustrated in (Figure 13.1, *left*). The first family is the *primary active transporters* (or *pumps*), which hydrolyze ATP in order to establish or maintain an electrochemical gradient for their substrates. These include P-type ATPases (Section 13.3.1.1) and ABC transporters (Section 13.3.1.2). Some active transporters can also use light energy for transport (Section 13.3.2).

The next family is the *secondary active transporters* (Section 13.4), which harness the energy of electrochemical gradients created by pumps to transport another substrate against its gradient. In other words, the secondary active transporters translocate two species, and couple the energetically-favorable transport of one species down its electrochemical gradient with the energetically-unfavorable transport of another against its gradient (Figure 13.1, *left*). These proteins are also known as *cotransporters* or *exchangers*. Transport is called *symport* if both substrates are moved in the same direction, and *antiport* if they are moved in opposite directions. A typical example of a symporter is the family of glucose transporters, which use a downhill gradient of sodium ions to transport

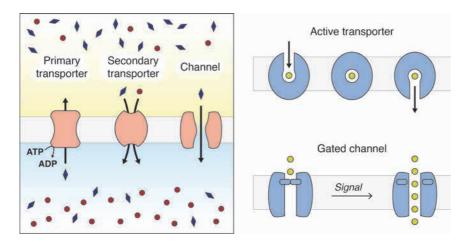


Fig. 13.1 ■ *Left*: Overview of the three different transport mechanisms. In this example, the secondary transporter is a symporter (rather than an antiporter), because both substrates are cotransported in the same direction. *Right*: The different mechanisms of active transporters and gated channels. Transporters possess an alternating-access mechanism and cycle between outward-open (*left*), occluded (*centre*), and inward-open states (*right*). Channels possess one continuous pathway when open, and their substrates passively move down their electrochemical gradients.

glucose against an uphill concentration gradient. A classical case of an antiporter is the sodium-calcium exchanger (NCX). While sodium is flowing down its gradient, calcium is transported the other way (counter-transported), out of the cell and against its own gradient. However, the net flow of ions is downhill.

Finally, the third family of transporters is the *channel proteins*, which passively permit the downhill movement of substrates (Figure 13.1, *left*). The transport rates in these proteins are limited by diffusion through the selectivity filter (the part of the protein that selects the correct substrate).

There is a very large difference in the time scales between the diffusion rates for channels and the large conformational changes that are necessary for substrate translocation in active transporters. The latter should never allow a continuous pathway between both sides of the membrane, because this would equilibrate the gradients and nullify the work and energy spent on transport — just like a bicycle pump that leaks the air pressure along the piston. Active transporters are therefore controlled by large conformational changes, which cycle the protein conformations in such a way that an inner cavity alternates between inwards- or outwards-facing forms (the *alternating access mechanism*; Figure 13.1, *right top*). As a consequence, the turnover rates of transporters are slow, on the order of milliseconds or more. On the other hand, channels, once open, provide a continuous pathway through the membrane and are limited only by the rate of diffusion. Turnover rates for channels are typically in the range of micro- to pico-seconds. This allows for swift responses and signal transmission. The voltage-sensitive channels involved in formation and locomotion of action potentials in nervous tissues are examples of this (Section 13.2.3.2).

In general, channels display selectivity for specific ions (such as Na⁺ over K⁺ or vice versa). Many channels are gated (Figure 13.1, right bottom), meaning that a signal either opens or closes them. Possible signals include cAMP, Ca²⁺, light, a change of membrane potential, or mechanical changes (see e.g. Section 13.2.3.2). Thus a gated channel is at the same time a signal transducer - or in other words, a receptor. Gated ion channels are denoted as ionotropic receptors. Similarly, conformational changes, such as those associated with substrate binding by the transporter, are exploited as a means of transmembrane signaling. A transporter system can be reduced into a non-transporting sensor system when a substrate binds to the protein and imposes a conformational change without subsequent transport. Intracellular factors can be recruited by the conformational change and activated for a downstream response. Examples of this kind include G-protein-coupled receptors (see Section 14.4.1), which appear to have evolved from bacteriorhodopsin into a highly diverse class of receptors. When these receptors control ion channels indirectly through second messengers, they are called metabotropic receptors, in contrast to the ionotropic receptors.

In Chapter 4, the basic structural features of membrane proteins were explored in detail. In the following sections, we shall see how a number of these structural features come together to form functional transporter proteins. Although it is not possible to describe here the structural features of all transporter folds and subfamilies determined thus far, the following examples together provide an overview of the major structurefunction principles in membrane transport.

Membrane Channels 13.2

13.2.1 β-barrel Porins

The porins are the archetypes of the β -barrel membrane proteins, which were discussed in detail in Section 4.7. Porins exist only in the outer membranes of gram-negative bacteria and eukaryotic mitochondria and plastids. The central cavity in the porins may be occupied or transiently covered by structural units that serve as a lid or plug in gating mechanisms. One example of this is in the bacterial OmpG protein responsible for oligosaccharide uptake (Figure 13.2).

Many porins do not bind their substrates with any significant affinity or specificity, but rather allow the passage of small polar substances (<600 Da) driven by their concentration gradients. However, for substances present at low concentrations passive diffusion is not sufficient. Such substances are therefore transported through substrate-specific channels, and examples include LamB, which is selective for maltose, PhoE, which is selective for phosphate and FecA, which imports ferric citrate.

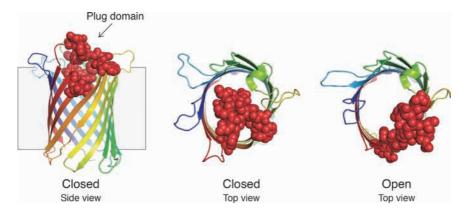


Fig. 13.2 ■ The bacterial outer membrane porin OmpG. The protein possesses a loop that acts as a lid, opening and closing the pathway across the membrane [PDB: 2IWW (closed form) and 2IWV (open form)].

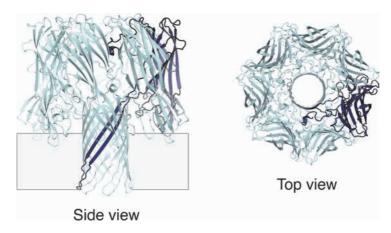


Fig. 13.3 ■ The α -hemolysin pore. The pore is formed from multiple subunits (one of which is colored dark blue) that assemble into a single β -barrel (PDB: 7AHL).

β-barrel proteins are also responsible for the pathogenicity of some bacteria and viruses, as they can facilitate invasion/spreading or even penetration of the host cell membrane. This can be achieved through large multimeric self-assembling β-barrel pores. A well-studied example is the bacterial α-hemolysin pore (Figure 13.3), which forms at the host cell membrane through oligomerization of the soluble α-hemolysin monomer. In fact, this mechanism resembles one of the key constituents of the human innate immune system, the membrane attack complex, which is formed from several complement factors (C5b, and C6, C7, C8 and C9) that assemble into a large transmembrane pore at the membranes of pathogens or other foreign cells (e.g. transplanted organs).

13.2.2 The Water Channel Aquaporin

Although water can diffuse through membranes, it is too inefficient to meet the physiological needs of many cells. Therefore, the existence of water channels was predicted long ago and eventually discovered in the early 1990s by Peter Agre. These channels are called aquaporins and are abundant proteins of many related types. Many cells possess such water-specific membrane channels, though a subset of proteins within the family are also permeable to glycerol, urea or ammonia. Aquaporin is a prime example of a difficult selectivity problem: how can the passage of small ions or protons be avoided (which would dissipate the hard-earned work of ion pumps), while permitting water to pass?

The aquaporins are tetramers, with independent pores through each subunit (Figure 13.4, left top). Each chain is composed of six transmembrane helices and two reentrant loops (Section 4.6.4.3) that "meet" at the center of the membrane (Figure 13.4, right). The N- and C-terminal halves of each subunit seem to have originated through a gene duplication and their structures are related by an approximate two-fold axis in the plane of the membrane (an "inverted repeat"; Section 4.6.6.2). The channel has an hourglass structure with two vestibules connected by a 20 Å long channel, which at its narrowest point is no more than 2.8 Å wide. Water molecules have to travel through this channel in single file (Figure 13.4).

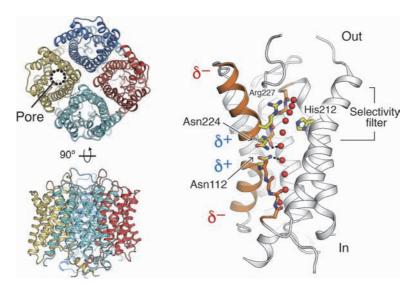


Fig. 13.4 ■ Left: The aquaporin tetramer has one pore in each chain (PDB: 3ZOJ). Right: One aquaporin monomer, showing the asparagines of the conserved NPA motifs (Asn112 and Asn224) and the arginine (Arg227) and histidine (His212) that restrict passage of H₃O⁺ ions. Reentrant loops are colored orange. Water molecules are shown as red spheres, and the hydrogen bonds between the asparagines and the central water molecules are indicated by black dashes. The selectivity filter backbone carbonyl groups from the reentrant loops are shown as orange sticks. Two helices have been removed for clarity.

The two reentrant loops contain a highly conserved signature motif, NPA, which is situated in the center of the membrane with the two NPA sequences juxtaposed. The asparagine residues from this motif (Asn112 and Asn224) hydrogen bond to water molecules in the middle of the channel (Figure 13.4).

The pore itself has a partly hydrophobic surface, but is also lined by several polar groups including carbonyl oxygens. This allows the polar water molecules to pass through the pore. The selectivity filter lies in the extracellular portion of the channel, and consists of polar backbone carbonyl groups and arginine, histidine, and phenylalanine side chains (Figure 13.4; phenylalanine not shown). The selectivity filter prevents positively charged ions such as H₃O⁺ from passing through the channel. This is achieved through (i) the positive charge of an arginine residue; (ii) the partial positive charges of the N-terminal helix dipoles of the two short helices in the reentrant loops; and (iii) the absence of a suitable number of ligands to compensate for the loss of the $\hat{H}_3\text{O}^+$ hydration shell. OH ions are excluded because they do not possess the ideal number of hydrogen bond donor groups to satisfy hydrogen-bonding requirements in the selectivity filter. Finally, an ultrahigh resolution structure of the protein (0.88 Å resolution) demonstrated that the line of water molecules in the channel do not share a continuous unidirectional hydrogen bonding network, hence preventing proton transport through the channel via the Grotthuss mechanism. In these ways, the aquaporin structure allows for exquisite selectivity of water over other small ions.

13.2.3 Potassium Channels

13.2.3.1 General structure of potassium channels

Potassium channels are found in both eukaryotes and prokaryotes. In eukaryotes, for example, they are found as voltage-gated channels in neurons. These channels close and open depending on the voltage, and they are of critical importance for the formation and modulation of action potentials in neurotransmission.

The K⁺ channels are members of a family of tetrameric proteins that also include channels for cations like Na⁺ and Ca²⁺. These proteins contain a variable number of transmembrane helices. The archetypical K⁺ channel is the bacterial protein KcsA, which has the simplest possible composition of only two transmembrane helices per protomer. The four subunits are arranged with a four-fold symmetry axis perpendicular to the membrane, creating an eight-helix bundle (Figure 13.5). However, unlike aquaporins (Section 13.2.2), the pore in K⁺ channels is formed through the *center* of the oligomer (Figure 13.5).

The KcsA tetramer possesses a selectivity filter, a central vestibule and an inner gating region (Figure 13.5). The selectivity filter is on the extracellular side of the protein, and is formed from the backbone regions of four reentrant loops (Section 4.6.4.3). The selectivity filter consists of a total of 16 carbonyl oxygens (four from each subunit) that point towards the center of the channel (Figure 13.5). The structure of the filter allows transport of K^+ ions but excludes the smaller Na^+ ions. This is because the distances between the

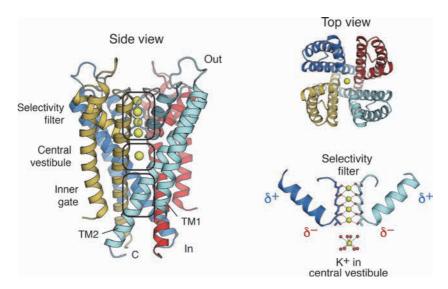


Fig. 13.5 ■ Structural features of the KcsA potassium channel. The tetramer possesses a selectivity filter, a central vestibule and an inner gate. Only two subunits are shown in the expanded view of the selectivity filter (right bottom). A number of K⁺ ions (yellow) fill the filter. Recent studies combining crystallographic analysis and molecular dynamics simulations indicate that at all positions are in fact occupied in the conducting channel. One ion is found in the central vestibule, coordinated by eight water molecules and stabilized by the negatively charged end of four helix dipoles (PDB: 1K4C).

carbonyl oxygen atoms in the filter are favorable for binding K⁺ ions, but not Na⁺ ions (which prefer shorter bond lengths). Hydrated potassium ions can therefore lose their shell of water molecules at a low energy penalty, and then travel along the filter through successive points of interaction with the four rings of carbonyl oxygens.

In contrast, dehydration of the smaller Na⁺ ions is highly energetically unfavorable since the coordination distances to the carbonyl oxygens of the filter are too long to compensate for loss of hydration. Therefore, Na⁺ ions will stay out of the selectivity filter and not pass through the channel. It has been disputed whether all four sites in the selectivity filter are occupied at the same time or if potassium only binds at two sites at alternating 1–3 or 2–4 positions (thus giving overall half occupancy at the four sites). Recent MD simulations and careful crystallographic analysis argue that the open, conducting channel has full occupancy at all four sites.

13.2.3.2 Gating of potassium channels

Potassium channels are gated, meaning that they are not open all of the time. Gating of potassium channels takes place on the inside face of the protein at a region known as the bundle crossing or the inner gate (Figures 13.5 and 13.6). The gate is formed from the C-terminus of the second TM helix of the pore (one helix from each of the four subunits). In the closed state of the channel the four helices associate with each other to prevent ion

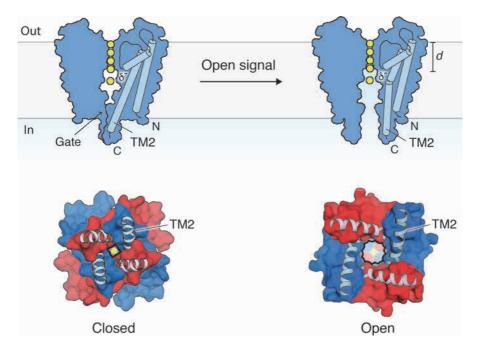


Fig. 13.6 ■ Structural changes during potassium channel gating. *Top*: Kinking of TM2 in response to a signal results in opening of the inner gate, thereby connecting the central vestibule with the cytoplasm. *Bottom*: A comparison between open and closed channel structures, viewed from the cytoplasmic face. The closed structure is of bacterial KcsA (PDB: 1K4C), and the open structure is the eukaryotic voltage-gated Kv1.2/2.1 channel (PDB: 2R9R, excluding voltage sensing domains).

flux. In response to a signal (e.g. voltage change, Ca^{2+} , pH), structural changes in the protein cause TM2 to become kinked at a conserved glycine or proline residue (Section 4.6.4.2). The C-termini of the helices consequently move laterally in the plane of the membrane, thereby opening the channel (Figure 13.6). This mechanism is sometimes described as being akin to the opening and closing of a camera iris. Due to the aqueous nature of the central vestibule and the helix dipoles from the selectivity filter (Figure 13.5), the effective transmembrane distance that the ions must travel, d, is significantly shorter than the true width of the membrane (Figure 13.6, *top right*).

One fascinating example of potassium channel gating lies in the voltage-gated potassium (Kv) channels, which are essential for action potential generation in neurons and cardiac cells. These proteins have an additional voltage-sensing domain (VSD) that precedes the pore-forming domain (Figure 13.7). The fourth TM helix of the VSD (S4) is rich in the positively charged residues arginine and lysine, which is very rare for a membrane-buried helix (Section 4.6.2). When the membrane potential is negative (more negative inside), the positively charged S4 helix is attracted towards the inside of the cell and the channel is closed. When the potential becomes positive, electrostatic forces on the S4 helix drive its movement towards the outside of the cell. This movement

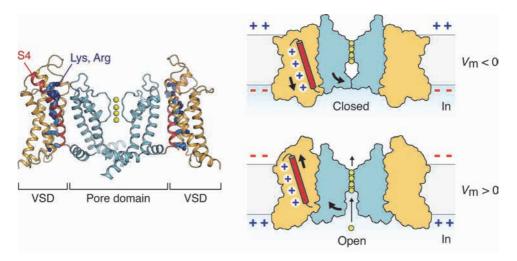


Fig. 13.7 ■ The mechanism of voltage-gated potassium channels. *Left*: Structure of the eukaryotic Kv1.2/2.1 channel. Only two voltage sensing domains (VSDs) and two subunits of the pore domain are shown. The positively charged residues on S4 are shown as blue sticks (PDB: 2R9R). Right: Schematic of the mechanism of voltage gating. The helix S4 in the VSDs moves up and down according to the membrane potential (Vm), opening and closing the entrance to the adjacent pore domain.

of S4 is transduced to the pore domain, resulting in opening of the pore and the flux of potassium out of the cell (Figure 13.7). The specific voltage at which the channel shifts from closed to open defines the voltage sensitivity of the channel.

In addition to voltage gating, the opening and closing of potassium channels can also be regulated by changes in pH (this is the case for KcsA), mechanical stress (mechanosensitive channels; see Section 6.2.1.1) or ligand binding. Examples of ligand-gated ion channels include the MthK channel activated by calcium binding, the inwardly rectifying channels (Kir) regulated by polyamines and cyclic-nucleotide (cAMP or cGMP) modulated channels. Ligand-induced gating is usually mediated by an additional cytosolic domain, to which the ligand can bind. Ligand binding induces a conformational change in the cytosolic domain, which is then transduced to the pore domain.

13.3**Primary Active Transporters**

Primary active transporters use an external energy source, such as light or ATP, to move solutes across membranes and against their electrochemical potential. For example, primary active transporters can pump H⁺ (as in bacteriorhodopsin; Section 13.3.2), cations like H⁺, Na⁺, K⁺ and Ca²⁺ (as in V- and P-type ATPase pumps; Section 13.3.1.1), and larger compounds (as in ABC transporters; 13.3.1.2).

13.3.1 ATP-Driven Transporters

13.3.1.1 P-type ATPases

The P-type ATPases a major class of primary active transporters that use the energy of ATP hydrolysis to transport solutes against their concentration gradients (the other class is the ABC transporters; Section 13.3.1.2). P-type ATPases account for approximately one-third of the ATP consumed by the body, in return generating and maintaining electrochemical gradients across cellular membranes — they charge the batteries, so to say.

P-type ATPases are found throughout the three domains of life, and almost all members of this family are cation transporters. Some important P-type ATPases are shown in Figure 13.8, and include the $\mathrm{Na^+/K^+}$ -ATPase (that maintains the plasma membrane electrochemical gradient), the $\mathrm{Ca^{2^+}}$ -ATPases (essential for intracellular $\mathrm{Ca^{2^+}}$ homeostasis and muscle relaxation), and the $\mathrm{H^+/K^+}$ -ATPase (that acidifies the stomach).

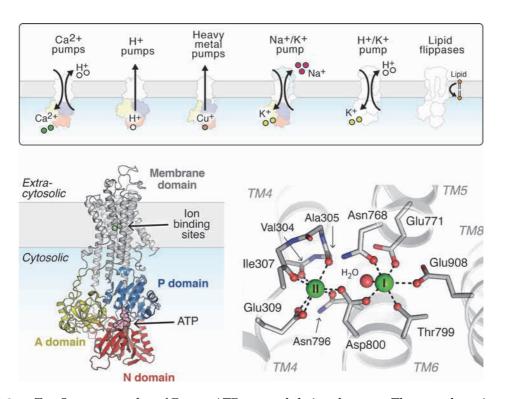


Fig. 13.8 ■ *Top*: Some examples of P-type ATPases and their substrates. The cytoplasm is colored blue. Pumps shown in color have had their structures determined. The lipid flippases transport lipids from the outer to the inner membrane leaflet. *Bottom left*: The overall architecture of P-type ATPases. Coordinates are from the Ca²⁺-transporting Sarco(endo)plasmic reticulum ATPase (SERCA, PDB: 1T5S). *Bottom right:* The Ca²⁺ binding sites from SERCA (PDB: 1T5T). Calcium ions I and II are colored green. Three backbone carbonyl groups are ligands to Ca-II through unwinding of TM4 within the membrane (i.e. a discontinuous helix; Section 4.6.4.3).

The conformational changes that mediate transport in the P-type ATPases are coupled to a cycle of phosphorylation and dephosphorylation of a conserved aspartate residue (hence the name "P-type" ATPase — as in "phosphorylation-type" ATPase). The phosphate (and energy) in this reaction comes from ATP, which is hydrolyzed by the protein to ADP and the terminal γ -phosphate transferred onto the aspartate residue.

All P-type ATPases share a common architecture that consists of three cytosolic domains and a membrane domain through which the ions are transported (Figure 13.8, bottom left). Some P-type ATPases, such as the Na^+/K^+ -ATPase, form a binary complex with a smaller (β) subunit.

All three cytosolic domains are essential for the ATP-dependent conformational changes in the membrane domain. The phosphorylation domain ("P-domain") is directly coupled to the membrane domain via two linkers. This domain houses the conserved Asp residue that is transiently phosphorylated during transport. The nucleotide-binding domain ("N-domain") is an insertion into the P-domain and it binds to the ATP nucleotide. Finally, the actuator domain ("A-domain") possesses a conserved sequence motif Thr-Gly-Glu-Ser (TGES) that catalyzes dephosphorylation of the P-domain.

The structures of numerous P-type ATPases have now been determined, though none is better studied than the Sarco(endo)plasmic reticulum Ca²⁺-transporting ATPase (SERCA). Structures of SERCA have been obtained in nearly all major conformational states along its reaction trajectory, providing us with a detailed understanding of the structural changes associated with ion transport. The pump cycles between so-called "E1" states, in which it has a high affinity for calcium, and "E2" states that have a low affinity for calcium but a high affinity for counter-transported protons. The major steps in the transport cycle are shown in Figure 13.9 and can be summarized as follows:

- (i) Two Ca²⁺ ions enter and bind to the transmembrane domain.
- (ii) and (iii) The protein rearranges to occlude the Ca^{2+} ions. ATP is hydrolyzed to ADP as the γ -phosphate is transferred onto the conserved Asp residue (Asp351 in SERCA).
- (iv) ADP leaves and the A-domain rotates, bringing the TGES motif close to the phosphorylated Asp residue. These movements cause opening of the membrane domain to the opposite side of the bilayer. Ca^{2+} leaves and 2–3 counter-transported H^+ ions bind.
- (v) A regulatory ATP binds to the N-domain. The TGES motif becomes fully engaged at the P-domain active site. The H⁺ ions are occluded. Glu from the TGES motif (Glu183 in SERCA) abstracts a proton from a water molecule, which then commences a nucleophilic attack on the phosphorylated Asp residue.
- (vi) The phosphate group leaves and the protein relaxes, opening an exit pathway into the cytosol through which the H^+ ions can diffuse. The protein then returns to state (i) and the cycle recommences.

Fig. 13.9 ■ The ATP-dependent conformational changes in the canonical P-type ATPase SERCA. The rotation of the A domain (about a vertical axis in the plane of the page) is indicated by a change in the position of the TGES motif.

13.3.1.2 ATP binding cassette (ABC) transporters

Like P-type ATPases, ABC transporters also utilize the energy of ATP hydrolysis to transport substrates across membranes. ABC transporters can function as either importers (in both prokaryotes and eukaryotes) or exporters (only in eukaryotes), and transport a wide variety of substrates such as ions, vitamins and even lipids. The ABC

transporter family includes numerous clinically-relevant proteins, most notably the CFTR protein involved in cystic fibrosis (which is an atypical member of this family) and the human p-glycoprotein which confers multidrug resistance in around half of human cancers.

ABC transporters share a common overall architecture: two transmembrane domains that accommodate and transport the substrate, and two cytosolic ATP binding cassette domains (hence the name "ABC" transporters) that bind and hydrolyze ATP (Figure 13.10, left). The transmembrane domains can be linked to the nucleotide-binding domains either as a single polypeptide chain, or they can be separate subunits that dimerize in the assembled transporter. The membrane domains from ABC transporters typically consist of two 6-10 helix bundles. The nucleotide-binding domains each possess half of a single ATP binding site, with the other half being provided by the opposing domain in a "head-to-tail" manner (Figure 13.10, right). Prokaryotic importers also possess periplasmic soluble substrate binding proteins, which associate with the substrate and deliver it directly to the membrane domain (Figure 13.10, left).

Similar to P-type ATPases (Section 13.3.1.1) and the secondary active transporters (Section 13.4), ABC transporters function via the principle of alternating access. The atomic structures of many full-length ABC transporters have been determined, allowing different possible mechanisms for solute transport to be postulated. The simplest and most widely accepted transport model is called the "Switch" or "Processive Clamp" mechanism, in which the nucleotide binding domains cycle through dimeric (ATP-bound)

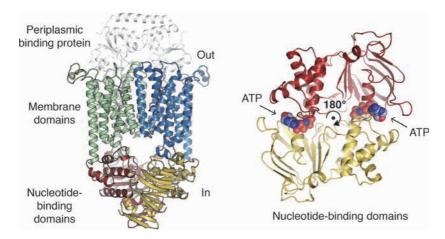


Fig. 13.10 ■ The overall structure of ABC transporters. *Left*: The structure of the vitamin D transporter BtuCD (colored cartoon), in complex with its periplasmic binding protein BtuF (white cartoon). In this example, the transporter is formed by four subunits: two membrane domains (green and blue) and two nucleotide-binding domains (red and yellow) (PDB: 4FI3). Right: ATPbound dimeric form of the nucleotide-binding domains from a multidrug transporter, viewed from below the membrane (PDB: 2ONJ).

Fig. 13.11 ■ The proposed switch mechanism of ABC transporters. A total of two ATP molecules are hydrolyzed per transport cycle.

and dissociated (nucleotide-free) states during ATP hydrolysis (Figure 13.11). The rigid-body movements of the nucleotide-binding domains are transmitted to the membrane domain, in turn causing alternating access of the substrate to each side of the membrane. The second model is the "Constant Contact" mechanism, in which the two nucleotide-binding domains do not fully dissociate from each other during the transport cycle, but still affect the structure of the membrane domains to mediate substrate translocation. The exact nature of the structural changes associated with substrate transport in ABC transporters are therefore yet to be determined.

13.3.2 Light-Driven Transport

Bacteriorhodopsin was the first ever membrane protein to have its structure determined at a near-atomic level (Section 4.3). It is a sophisticated yet simple proton pump system in archaea that utilizes the energy of sunlight to create a proton gradient across the cell membrane (Figure 13.12). The energy of this proton gradient can then be exploited by the protein ATP synthase (Section 8.3.2) to generate ATP for cellular use.

Bacteriorhodopsin exists as a homotrimer in the native membrane, with each subunit functioning as a proton transporter. Critical to the mechanism of bacteriorhodopsin is a retinal chromophore (one per subunit), which absorbs the energy of light. It is covalently linked to the protein via a Schiff's-base linkage to Lys216 (Figure 13.12), and photon absorption stimulates isomerization of the retinal from an all-*trans* ground state to a 13-cis high-energy state. Due to the covalent link to Lys216, the photoisomerization causes reconfiguration of the protein that effectively separates the Schiff's base from its ionic

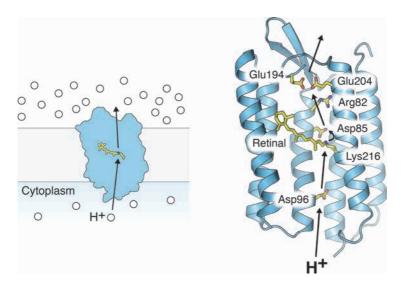


Fig. 13.12 ■ The structure of bacteriorhodopsin. Arrows indicate the proton translocation pathway. While the protein is a homotrimer in the native membrane, only the monomer is shown for simplicity (PDB: 1C3W).

interaction with Asp85 (Figure 13.12). Isolated charges in the membrane are very unfavorable, and this causes an increase in the pKa value for Asp85. Because of this, a proton is transferred from the Schiff's base to neutralize the Asp85 side chain.

During relaxation of the high-energy state, Arg82 approaches Asp85 and stimulates its deprotonation, resetting the pKa. The proton escapes to the extracellular environment via a Glu194–Glu204 pair. The Schiff's base is reprotonated from the cytoplasmic side via a buried Asp96 residue. Therefore, this process creates a proton-transporting chain that pumps one proton from the cytoplasm to the extracellular environment per photoisomerization cycle. This mechanism — a proton-conducting pathway including titratable Asp and Glu residues modulated in pKa value by an Arg residue — occurs in many proton-pumping systems.

13.4 Secondary Active Transporters

Secondary active transporters belong to the solute carrier superfamily of membrane proteins (SLCs). This superfamily contains almost 400 members that are currently classified into subfamilies SLC1 through 52.

13.4.1 Antiporters

13.4.1.1 Cation/proton antiporters (CPA): NhaA

The cation/proton antiporters (CPA, also classified as the SLC9 family) are a superfamily of transporters found in all organisms. These proteins translocate specific cations (primarily Na⁺) in exchange for protons, thereby helping to regulate intracellular pH, Na⁺ levels and cellular volume.

The best-characterized member of the CPA superfamily is the $E.\ coli\ Na^+/H^+$ exchanger, NhaA. The function of NhaA is to regulate intracellular concentrations of Na⁺ and H⁺, and allow $E.\ coli$ to survive in high salt conditions and alkaline pH. However, the protein can also act as a Li⁺/H⁺ exchanger under conditions of Li⁺ toxicity.

NhaA sits in the inner membrane and transports Na^+ ions out of the cytosol using the energy derived from the proton gradient (one Na^+ in exchange for two H^+ ions per transport cycle). The activity of NhaA is highly dependent on the pH of the cytoplasm, being inactive at ~pH < 6.5, and fully active at ~pH > 8. This pH dependence prevents the accumulation of a high concentration of H^+ ions in the cytoplasm.

NhaA is a homodimer and its overall architecture is shown in Figure 13.13. Ions are transported through both of the two subunits, which consist of 12 transmembrane helices each. Each monomer is made up of a dimerization domain (TM1–2 and 8–9) and a "core" domain (TM3–5 and 10–12), through which the ions are transported. The final two helices, TM6–7, contribute to the stability of the protein.

Like many other secondary active transporters, such as LeuT (Section 13.4.2.1) and EmrE (Section 13.4.1.2), NhaA possess an inverted repeat topology (Section 4.6.6.2). The inverted repeat is readily evident in the core domain, where TM3–5 and TM10–12 are

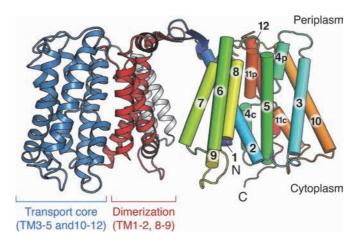


Fig. 13.13 ■ Structure of the NhaA homodimer. One chain is shown in ribbon representation and colored according to domain. The other is shown as cylinders and colored in rainbow from N- to C-termini (PDB: 4AU5).

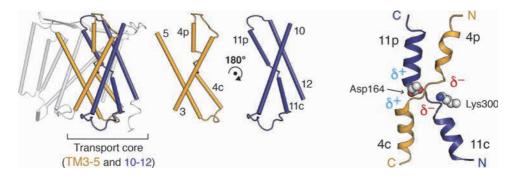


Fig. 13.14 ■ *Left*: Inverted repeat in the NhaA core domain. The two-fold rotational symmetry axis goes into the page (represented by a black dot) and lies parallel to the plane of the membrane. Only the monomer is shown. Right: Hallmark crossed discontinuous helices in NhaA. The discontinuous helices TM4 and TM11, where suffixes "p" and "c" indicate the periplasmic and the cytoplasmic halves, respectively (PDB: 4AU5).

related to each other by a two-fold axis in the plane of the membrane (Figure 13.14, left). The dimerization domain is also an approximate inverted repeat formed from TM1-2 and TM8-9.

The presumed ion binding site in NhaA involves two discontinuous helices (Section 4.6.4.3 and Figure 13.14, right). Although this is common in membrane transporters such as P-type ATPases (Section 13.3.1.1) and LeuT (Section 13.4.2.1) — the discontinuous helices in NhaA and the wider CPA superfamily form a unique structural motif and cross over each other in the center of the membrane. This places the like charges from the opposing helical dipoles in close proximity to each other. These helical dipoles are neutralized by the side chains from Asp164 and Lys300 (Figure 13.14, right), which are critical for the activity of the protein.

To date, the structure of NhaA has only been determined in the H+-bound inwardopen inactive conformation. This is because crystals were obtained at pH \leq 4 at which the protein is not active. Therefore, the sodium binding site and the outward-facing protein conformation have not yet been directly visualized. Nonetheless, a combination of structural analysis, sequence analysis, mutational studies, molecular dynamics simulations and electrophysiology has identified the probable ion binding site(s), and allowed a possible transport mechanism to be postulated.

NhaA functions via the alternating-access mechanism (Section 13.1), cycling between inward-open and outward-open states. There is a negatively charged pathway in the inward-open conformation between the core and the dimerization domains, which forms the cytoplasmic entry and exit pathway for the transported ions. At the end of the tunnel lie the discontinuous helices and the conserved residues Asp164, Lys300 and Asp163 (Figure 13.14, left; Asp163 is omitted for clarity). The Na⁺ binding site is likely to be formed from both Asp163 and Asp164. The acidic side chain of Asp164 is also the binding

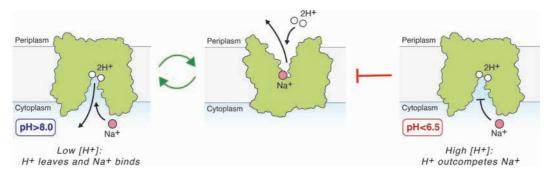


Fig. 13.15 ■ Ion transport by NhaA, regulated by cytoplasmic pH. Na⁺ is transported against its concentration gradient and H⁺ down its gradient. For simplicity, only one chain of the NhaA dimer is shown.

site for one of the counter-transported H⁺ ions, while the other H⁺ most likely binds at one of either Asp163 or Lys300.

It has now been shown that the pH dependence of the protein arises from competition between Na⁺ and H⁺ for the ion binding sites, which is a consequence of having one (Asp164) or more (possibly also Asp163) common ligands for the two ions. When the pH of the cytoplasm is low, the high concentration of H⁺ means that the side chain of Asp164 (and also the second H⁺ binding site) will be protonated. This means that Na⁺ cannot bind and the protein is inactive (Figure 13.15, far right). However, when the pH of the cytoplasm is raised, Asp164 (and also the second H⁺ binding site) becomes deprotonated, and its charged -COO side chain can then participate in Na binding. Na binding triggers a conformational rearrangement in which the protein adopts its outward-open form, allowing Na⁺ to leave and H⁺ ions to bind and be transported down their concentration gradient.

13.4.1.2 Small multidrug resistance (SMR) transporters: EmrE

The small multidrug resistance (SMR) transporters are bacterial inner-membrane proteins that confer resistance to numerous cytotoxic compounds. These proteins are the smallest of the secondary active transporters, functioning as homo- or hetero-dimers that comprise only ~120 amino acids per monomer. The SMR transporters are exemplified by the E. coli protein EmrE. EmrE is an antiporter (Section 13.1), and exploits the protonmotive force to couple the movement of protons down their concentration gradient with the translocation of polyaromatic cations (e.g. dequalinium, acriflavine and tetraphenylphosponium) out of the cell. The protein transports two protons for each molecule of exported substrate.

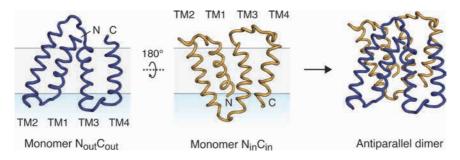


Fig. 13.16 ■ The dual topology of EmrE. The protein forms an antiparallel homodimer with a pseudo-two-fold axis of symmetry in the plane of the membrane (PDB: 3B5D).

EmrE is a fascinating structural example of a secondary active transporter. In Sections 4.6.6.1 and 4.6.7.2, we saw how helical membrane proteins have a preferred topology that is encoded in their sequences and established during insertion into the bilayer. However, structural and biophysical studies have established that EmrE represents a notable exception to this rule. EmrE does not possess a clear charge bias between "inside" and "outside" loops that would dictate its orientation according to the "positive-inside rule" (Section 4.6.6.1). With the absence of other strong topogenic signals (such as a large cytoplasmic N-terminal domain), the protein therefore inserts with equal probability in $N_{in}C_{in}$ and $N_{out}C_{out}$ orientations (Figure 13.16). For this reason, EmrE is known as a dual-topology protein.

It was long debated whether EmrE dimerized in a parallel or antiparallel fashion, but overwhelming data eventually determined that the physiologically relevant form is the antiparallel state (Figure 13.16). The formation of "head-to-tail" homodimers creates an approximate two-fold symmetry axis in the plane of the membrane, a feature that was discussed in Section 4.6.6.2 and is observed in many membrane transporters. However, while most secondary active transporters possess internal structural symmetry within a single polypeptide chain (e.g. NhaA in Section 13.4.1.1), in EmrE the symmetry occurs between two identical subunits arranged in an antiparallel way.

The symmetry of EmrE is integral to its transport mechanism. The protein transports its substrates via the principle of alternating-access (Section 13.1 and Figure 13.17). Because of its pseudosymmetry, the overall conformation of the protein will be the same in each of these two states (since the two subunits interchange their conformations). The direct coupling of proton transport to substrate extrusion is achieved though the existence of a single binding site in the center of the protein. This site is formed by two negativelycharged glutamate residues (one per monomer), and each of the two substrates competes for binding. Therefore, when the substrate-loaded protein is open to the periplasm where the proton concentration is high, the substrate will be replaced by two protons. These two protons are then transported down their concentration gradient into the cell, where they dissociate and the cycle starts anew (Figure 13.17).

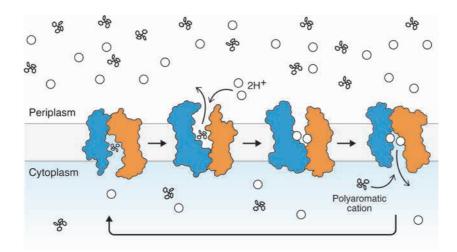


Fig. 13.17 ■ Overview of substrate transport by EmrE. Protons (white circles) are transported down their concentration gradient and polyaromatic cations (here tetraphenylphosphonium) are transported against their concentration gradient. During the transport cycle, the two subunits exchange conformations as the protein switches between inward- and outward-facing states.

13.4.1.3 Resistance-nodulation-division (RND) transporters: AcrB

Gram-negative bacteria possess both inner and outer cell membranes. This means that the extrusion of substances from the cellular environment requires transport across both. For many compounds or ions it will be enough to transport them across the tight, inner membrane with release to the periplasmic space, where spontaneous diffusion through pores of the outer membrane can then take care of the rest. However, for hydrophobic substances that spontaneously repartition into the inner membrane, or ions/substances that bind tightly to any protein, this mechanism would not suffice. This is because these substances would either return to the cell or damage the important proteins exposed to the periplasmic space. Therefore, in gram-negative bacteria, the Resistance-Nodulation-Division (RND) transporters have evolved to perform coupled transport across *both* membranes along a single, connected pathway that leads directly to the extracellular environment.

The RND transporters perform active transport coupled to the proton-motive force across the inner membrane. They transport a broad range of substrates, such as many different antibiotics (therefore contributing to antibiotic resistance) and toxic copper, yet discriminate against vital compounds like glucose and amino acids. Important RND efflux systems include AcrAB-TolC and MexAB-OprM (multidrug resistance determinants) and CusABC (toxic copper extrusion). Our knowledge of the structure and mechanism of action of RND transporters is based on a combination of what we have learned from each of these three homologous systems, primarily via X-ray crystallography of

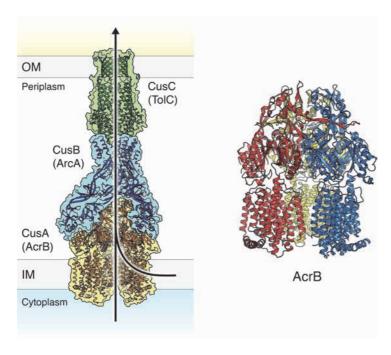


Fig. 13.18 ■ Structural overview of AcrB and the CusABC (AcrAB-TolC) efflux complex. Left: The structure of AcrB. The functional protein is an asymmetric homotrimer (PDB: 1IWG). Right: Schematic of the bacterial CusABC copper efflux pump. The CusABC complex is homologous to the AcrAB-TolC multidrug efflux pump. "OM" and "IM" denote the outer and inner membranes, respectively. [PDB: 3PIK (CusC) and 3NE5 (CusAB)].

individual components and electron microscopy of holocomplexes. RND transporters typically consist of three subunits: (i) the AcrB, MexB or CusA proton-coupled active antiporter unit residing in the inner-membrane, (ii) the periplasmic AcrA, MexA or CusB adaptor unit and (iii) the TolC, OprM or CusC outer membrane porin unit (Figure 13.18, left). Each of these subunits is arranged as a homotrimer, except for the AcrA/MexA/ CusB adaptor unit, which can also be a hexamer or a nonamer.

The critical input comes from the proton-coupled antiporter, which ensures the substrate loading and active transport along the RND pathway (Figure 13.18, right). This subunit has been studied in great detail for AcrB. The first structures of AcrB showed a symmetric, empty trimer arranged in a trigonal crystal form. This crystal form has been inadvertently reproduced by numerous laboratories that were attempting to crystallize entirely different proteins. This is because AcrB is often upregulated under antibiotic selection used in laboratory. Furthermore, it binds to a Ni²⁺ affinity column used for purification and is very easily crystallized even under very scarce concentrations as a membrane protein contaminant. However, the AcrB trimer can also crystallize as an asymmetric trimer showing an access state a binding state and an empty extrusion state for transported substrate. The access and binding state show peripheral or deep binding,

respectively, to the substrate binding site exposed to the membrane. In contrast, the extrusion state shows binding site occlusion to the outside and opening towards the inner transport pathway. The asymmetric trimer is furthermore connected to the protonation pathway that supports proton antiport some 50 Å away from the substrate binding site. Conformational changes associated with protonated/deprotonated states of key Glu/Asp residues and how they interact with conserved Lys/Arg residues affect the substrate binding sites and control the sequential interchange of the three states within the trimer.

13.4.2 Symporters

13.4.2.1 Neurotransmitter: sodium symporters (NSS): LeuT

Different symporter families transport amino acids and related compounds into the cell using the Na⁺ gradient. In animals, amino acid transporter families are also responsible for the reuptake of neurotransmitters from synaptic clefts. One such class of transporters (SLC6) is denoted the neurotransmitter sodium symporter (NSS) family, which perform the active reuptake of neurotransmitters and/or amino acids like dopamine, serotonin, noradrenaline, γ-aminobutyric acid (GABA), glycine and non-polar amino acids. These transporters typically have 11 or 12 transmembrane helices and cotransport two Na⁺ ions together with their substrate. Many mammalian transporters of the family also cotransport Cl⁻, and some counter-transport K⁺. For example, the serotonin transporter (SERT) transports one Na⁺, Cl⁻ and serotonin into the cell with counter-transport of one K⁺. Malfunction of neurotransmitter transporters can lead to a number of disorders such as epilepsy, Parkinson's disease, depression and autism. On the other hand, inhibition of these transporters by psychiatric drugs can favorably prolong the activity of neurotransmitters at the synaptic cleft to help treat psychiatric disorders. The serotonin transporter, for example, is inhibited by "selective serotonin reuptake inhibitors" (SSRI drugs), which include antidepressants such as fluoxetine (Prozac) and citalopram (Cipramil). The tricyclic antidepressants (TCA) such as clomipramine represent another very important class of inhibitors of clinical use, and are slightly less specific to their molecular target.

Amino acid transporters with high sequence identity to human NSS proteins are also found in many bacteria and archaea. The first structure from the NSS family came with the amino acid transporter LeuT from the bacterium *Aquifex aeolicus*. Later, the structures of the *Drosophila* dopamine transporter dDAT and the *Bacillus halodurans* multiple hydrophobic amino acid transporter MhsT were also determined. Structures of these NSS proteins have been obtained in several states of the functional cycle, revealing also key aspects of how the transport cycle is regulated and inhibited.

LeuT transports a range of amino acids such as leucine and alanine. The protein is a homodimer, and each chain consists of 12 transmembrane helices (Figure 13.19, *left top*) that are arranged in a 5+5 two-fold inverted topology (Section 4.6.6.2). That is, TM1–5 are

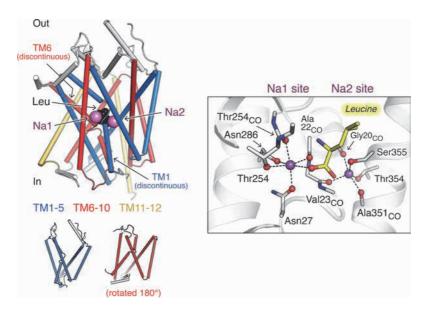


Fig. 13.19 ■ Structural features of LeuT. Left top: The structure of LeuT in substrate-bound outward-occluded state. Leucine, the transported substrate, is shown as black spheres. The co-transported Na⁺ ions are purple (PDB: 3F3E). Left bottom: The internal structural symmetry between TM1-5 and TM6-10 is evident. To facilitate comparison, TM6-10 has been rotated in the figure. Right: The Na⁺ binding sites. Subscript "CO" indicates that the bonds arise from the backbone carbonyl groups. Ala22_{CO} and Val23_{CO} arise from the unwound portion of TM1, highlighting the importance of the discontinuous helices.

related to TM6-10 through a pseudo-two-fold axis in the plane of the membrane (Figure 13.19, left bottom). Some additional helices are found in loops on both sides of the membrane and two additional, C-terminal segments, 11 and 12, form the dimerization interface. Core helices TM1 and TM6 are discontinuous in the middle of the membrane (Section 4.6.4.3), surrounding the substrate binding site. The unwound helices are critical for formation of the substrate and Na⁺ binding sites. The Na⁺ binding site 1 (Na1) is near the substrate site, and for transported amino acids the carboxylate group is a direct ligand to Na1 (Figure 13.9, right). Na⁺ binding site 2 (Na2) is special, showing trigonal bipyramidal coordination in outward-oriented states (rather than octahedral-like Na1). Two of the five ligands are main chain carbonyl oxygens (Gly20_{CO} and Val23_{CO}) from the N-terminal part of transmembrane helix 1 (TM1a; Figure 13.9, right).

In the outward-oriented state of LeuT, a large hydrophobic cavity leads to the extracellular environment as a substrate entry pathway, while a thin gate including an Arg-Asp salt bridge occludes the bound substrate (Figure 13.20). Substrate binding can only take place along with Na⁺, and only with both Na⁺ ions and an amino acid substrate bound can the transporter reach this occluded state. Once occluded, the transporter can now switch from the outward to the inward-oriented state. The inward-oriented state

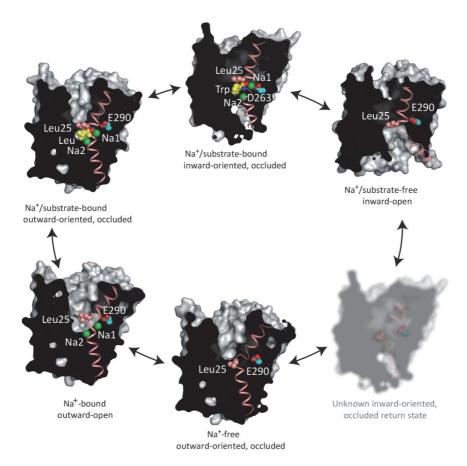


Fig. 13.20 ■ The transport mechanism of LeuT, an amino acid transporter of the NSS family, showing outward and inward occluded states for Na⁺/substrate uptake and H⁺ counter-transport. Figure kindly provided by Dr. Lina Malinauskaite.

is reached by closure of the extracellular, hydrophobic cavity, where at the same time the transmembrane segment 5 (TM5) is extended and deformed at the intracellular interface. This structural change at TM5 allows the intracellular environment to reach the Na2 site and add a cytoplasmic water molecule as a sixth ligand in octahedral coordination at Na2 (previously trigonal bipyramidal in the outward-oriented state). This facilitates Na⁺ release, which might also be stimulated by a negative membrane potential. Once Na2 is released, the tight coordination of TM1a is lost, and it opens up to allow further release of Na1 and the bound substrate to the cytoplasm. Return of the transporter along with H⁺ counter-transport (in the case of amino acid transporters) is associated with reorientation of a conserved Leu residue of the unwound core segment of TM1 occupying the empty transport site.

For a transporter to work efficiently, it must be dynamic yet also prevent uncoupled activity. One way to ensure this is to lower the energy barrier between the outward- and

inward-facing occluded state by a combination of energy penalties and gains associated with each of the occluded states associated with the transition. At the same time, only the combined Na⁺ and substrate bound form is occluded in the outward-facing state and can enter this smooth energy landscape. Uncoupled Na⁺ dissipation or downstream substrate efflux are therefore prevented while a dynamic transport function is facilitated.

Inhibitors can interfere either with the extracellular cavity (blocking it from closure and hence transition to the inward-facing states), or by deeper binding to the substrate site (also blocking occlusion). LeuT crystallized with tricyclic antidepressants showed binding at the extracellular cavity and thus stabilization of the outward-oriented occluded state (Figure 13.20). In contrast, dDAT structures with similar compounds showed, as expected, that they bind at the substrate binding site.

For Further Reading

Original Articles

- Doyle DA, Morais Cabral J, Pfuetzner RA, *et al.* (1998) The structure of the potassium channel: Molecular basis of K+ conduction and selectivity. *Science* **280**: 69–77.
- Grigorieff N, Ceska TA, Downing KH, *et al.* (1996) Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J Mol Biol* **259**: 393-421.
- de Grotthuss CJT. (1806), translated from French as: Memoir on the decomposition of water and of the bodies that it holds in solution by means of galvanic electricity. (2006) *Biochim et Biophys Acta Bioenergetics* **1757**: 871–875.
- Hunte C, Screpanti E, Venturi M, et al. (2005) Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH. *Nature* **435**: 1197–1202.
- Jiang Y, Lee A, Chen J, et al. (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**: 515–522.
- Korkhov VM, Mireku SA, Locher KP. (2012) Structure of AMP-PNP-bound vitamin B12 transporter BtuCD-F. *Nature* **490**: 367–372.
- Kosinska-Eriksson U, Fischer G, Friemann R, et al. (2013) Subangstrom resolution X-ray structure details aquaporin-water interactions. *Science* **340**: 1346–1349.
- Lee C, Yashiro S, Dotson DL, et al. (2014) Crystal structure of the sodium-proton antiporter NhaA dimer and new mechanistic insights. *J Gen Physiol* **144**: 529–544.
- Long SB, Campbell EB, Mackinnon R. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. *Science* **309**: 897–903.
- Long SB, Tao X, Campbell EB, MacKinnon R. (2007) Atomic structure of a voltage-dependent K+ channel in a lipid membrane-like environment. *Nature* **450**: 376–382.
- Malinauskaite L, Quick M, Reinhard L, et al. (2014) A mechanism for intracellular release of Na+ by neurotransmitter: Sodium symporters. *Nature Struct Mol Biol* **21**: 1006–1012.

- Malinauskaite L, Said S, Sahin C, *et al.* (2016) A conserved leucine residue occupies the empty substrate site in a return state of the neurotransmitter: Sodium symporter LeuT. *Nature Comm* 7: 11673.
- Morrison EA, DeKoster GT, Dutta S, et al. (2012) Antiparallel EmrE exports drugs by exchanging between asymmetric structures. *Nature* **481**: 45–50.
- Olesen C, Picard M, Winther AM, et al. (2007) The structural basis of calcium transport by the calcium pump. *Nature* **450**: 1036–1042.
- Song L, Hobaugh MR, Shustak C, et al. (1996) Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **274**: 1859–1866.
- Toyoshima C, Nakasako M, Nomura H, Ogawa H. (2000) Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405**: 647–655.
- Walz T, Hirai T, Murata K, *et al.* (1997) The three-dimensional structure of aquaporin-1. *Nature* **387**: 624–627.
- Winther AM, Bublitz M, Karlsen JL, et al. (2013) The sarcolipin-bound calcium pump stabilizes calcium sites exposed to the cytoplasm. *Nature* **495**: 265–269.
- Yildiz O, Vinothkumar KR, Goswami P, Kuhlbrandt W. (2006) Structure of the monomeric outermembrane porin OmpG in the open and closed conformation. *Embo J* **25**: 3702–3713.
- Zhou Y, Morais-Cabral JH, Kaufman A, MacKinnon R. (2001) Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 Å resolution. *Nature* **414**: 43–48.

Reviews

- George AM, Jones PM. (2012) Perspectives on the structure-function of ABC transporters: The switch and constant contact models. *Prog Biophys Mol Biol* **109**: 95–107.
- Gouaux E, Mackinnon R. (2005) Principles of selective ion transport in channels and pumps. *Science* **310**: 1461–1465.
- Henzler-Wildman K. (2012) Analyzing conformational changes in the transport cycle of EmrE. *Curr Opin Struct Biol* **22**: 38–43.
- Møller JV, Olesen C, Winther AM, Nissen P (2010). The sarcoplasmic Ca2+-ATPase: Design of a perfect chemi-osmotic pump. *Q Rev Biophys* **43**: 501–66.
- Schlessinger A. (2013) SLC classification: An update. Clin Pharmacol Ther 94:19-23.
- Yan N. (2015) Structural biology of the major facilitator superfamily transporters. *Ann Rev Biophys* **44**: 257–283.

Signal Transduction

Signal transduction is a central topic in all organisms and this is particularly true for multicellular organisms. In principle, all activities need to be regulated. This can be done by smaller signaling molecules or by macromolecules. Several chapters have already touched on different signaling systems. Transcription (Chapter 10) is highly regulated and so is the degradation of macromolecules (Chapter 12). One main route of communication of external signals to the nucleus is by protein phosphorylation. The immune system (Chapter 17) is also an example of complex signaling. The signaling often occurs as a cascade, where a chain of sequential reactions leads to the final effector.

14.1 Signaling Controls the Cellular Activity from the Outside

Many of the activities in the eukaryotic cell are controlled by factors outside the cell. This is especially true in multicellular organisms, where different kinds of cells in various kinds of tissues need to respond differently to make the whole organism function properly.

Outside factors influence the cells through *signaling pathways*. In most pathways, the factor binds to a receptor molecule in the cell membrane. The receptor molecule transfers the signal (but not the molecule) to the inside of the membrane, where other molecules react to the signal, multiplying the effect, and at the end this leads to the desired activity in the cell. The factor may be a small molecule like epinephrine or odorants, a short peptide-like glucagon, or proteins like growth hormone and the interferons. Some pathways use lipophilic factors that can pass the cell membrane and activate intracellular receptors. The whole process is called signal transduction.

There are two main types of signaling pathways:

- (i) Some signals require immediate changes in the state of the cell, for example, when the cells of our visual system react to photons. In this case, the signal leads to a quick change in the state of the cell that can be sensed by nerve cells and create an impulse to the brain. The signal leads to a change in activity of existing proteins. Signaling pathways of this kind often use G-protein coupled receptors and trimeric G-proteins that can activate various effector molecules.
- (ii) Other signals lead to a more lasting change of the property of the cell and changes in the amount of specific proteins. Here, the signaling pathway ends with the activation or deactivation of a transcriptional repressor in the nucleus. The outside signal in these signaling pathways is often a hormone or a growth factor produced by other cells.

Inside the cells, phosphorylation and dephosphorylation of proteins (Section 14.2.2) are two of the main methods to control various activities. Many signaling pathways involve protein kinases — enzymes that phosphorylate the hydroxyl group of tyrosine or serine/threonine residues, and these kinases are in turn regulated by phosphorylations and dephosphorylations of tyrosine and serine/threonine residues of the kinases themselves. Specific phosphatases remove the phosphate groups to revert the effect of the kinases, and these enzymes can also be regulated by extracellular signals.

The properties of G-proteins (Section 8.3.3) are used in many signaling pathways. They act as molecular switches on the inside of cells. G-proteins are used in signaling due to their capacity to be switched to the active (ON) state. The lifetime of this state is controlled by receptors and other proteins involved in the signaling pathway. They can switch the G-protein to the OFF state.

At least two important groups of G-proteins are involved in signaling: the trimeric G-proteins linked to the G-protein coupled receptors and a large group of monomeric G-proteins, the Ras superfamily (Table 14.1).

How is the signal communicated through the cell membrane? All cell membrane receptors have to convert binding of an extracellular ligand to a signal on the inside of the membrane. Changes on the extracellular side of the membrane influence the conformation

TABLE 14.1 Families of Monomeric G-Proteins in the Ras Super-**Family and Their Functions**

| Ras | Cytoplasmic signaling pathways controlling gene expression |
|-----|--|
| Rho | Regulation of cytoskeletal growth |
| Rab | Regulation of intracellular vesicular transport |
| Ran | Nuclear import and export |
| Arf | Regulation of vesicular transport |

or arrangement of the cytoplasmic domains on the inside of the membrane. There are several ways to do this:

- conformational changes of oligomeric receptors;
- conformational changes in monomeric receptors;
- channels.

To illustrate some mechanisms in signaling pathways, a few examples will be discussed: the cytokine receptor pathways, pathways using receptor tyrosine kinases and pathways using G-protein coupled receptors.

Signaling by Cytokines 14.2

The cytokines are small proteins (5–20 kDa) with important roles in signaling. They are released from cells and affect the behavior of other cells. They circulate at very low concentrations, which can be significantly increased at special instances. There are several types of cytokines, like interleukins and interferons. One large group of cytokines belongs to the four α -helix bundle family.

The cytokine receptors form a large group of receptor molecules that bind cytokines and some small hormones. The binding leads to a signal that directs cells to differentiate, divide or undergo apoptosis. These receptors are membrane proteins with an extracellular part, a single transmembrane segment, and an intracellular domain. In some cases, for example, the growth hormone receptor, two identical receptor molecules are involved, but in other cases two different receptor molecules (often called the α - and the β -subunits) form the active complex with the cytokine. The extracellular part has a receptor domain and various other domains depending on the receptor family.

The cytokine receptors are usually dimers, which by binding the cytokine leads to the signaling. This activates tyrosine kinases of the Janus kinase family (JAK). These large kinases phosphorylate the cytoplasmic domain of the receptor and STAT (signal transducers and activators of transcription) molecules are attracted to the phosphorylated receptor and, through further steps, this leads to control of transcription in the nucleus (the JAK-STAT signaling pathway). There are seven members of the STAT family, each with a range of genes that they regulate.

14.2.1 Signaling Through the Growth Hormone-Receptor

Signaling using growth hormone (GH) has been of significant interest for a long time. In 1992, a crystal structure was determined for the complex between the extracellular part

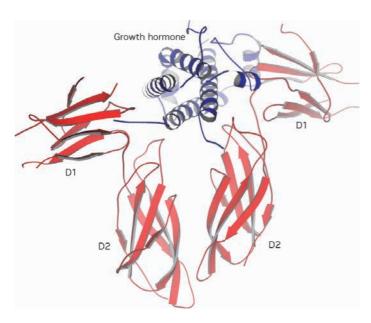


Fig. 14.1 ■ The extracellular part of the GH receptor with the bound hormone. The hormone, like many cytokines, is a four-helix bundle. The two identical monomers of the receptor have two domains with the fibronectin III fold. The connections to the transmembrane helices and the intracellular parts are at the bottom of the drawing (PDB: 3HHR). The binding of the hormone leads to a conformational change of the intracellular part of the receptor.

of the GH receptor and its ligand growth hormone. The extracellular part of the GH receptor has two domains (D1, D2) with the fibronectin type III fold (Figure 14.1). This type of fold is very common in extracellular parts of receptor molecules (Table 14.2) and very similar to the Ig-fold (Chapter 17). The intracellular domain of the growth hormone receptor has about 350 amino acid residues. Surprisingly, the complex showed that a single hormone molecule binds to two receptor molecules (a 1:2 complex). Loops at one end of each of the receptor domains form the binding surface. The ligand-binding loops of the D1 domain correspond to the hypervariable loops in immunoglobulins, but the D2 domain exposes the opposite end of its β sandwich to the ligand. The model for the activation has been that the receptor monomers dimerize due to the hormone binding. More recent experiments have given a different view. The receptor without ligand is predominantly a dimer held together by the transmembrane helices. Hormone binding leads to a "scissor-like" movement of the receptor (Figure 14.2).

The hormone molecule is asymmetric, but the binding elements in the two receptor monomers are the same. Because of the asymmetry of the ligand, the receptor dimer also becomes asymmetric. The affinity of two macromolecules binding to each other is roughly proportional to the area of the binding surface. The interacting surface is considerably greater in one of the receptor-ligand interfaces. This suggests a mechanism for the ligand-induced conformational change of the receptor. In the presence of the hormone, one

TABLE 14.2 Examples of Domains Found in Extracellular Portions of Receptors and Other **Extracellular Proteins**

| Domain Name | Size and Conformation | Examples of Proteins Containing the Domain |
|--|---|---|
| Fn1 (Fibronectin type 1) | ~40 residues, two small sheets linked with disulfide bonds | Fibronectin, extracellular binding proteins |
| Fn2 (Fibronectin type 2) | ~40 residues, two small sheets linked with disulfide bonds | Fibronectin, proteins in blood coagulation |
| Fn3 (Fibronectin type 3) | ~100 residues, antiparallel β sandwich, very similar to Ig | Fibronectin and other surface binding proteins, cytokine receptors, receptor tyrosine kinases |
| Receptor L domain | ~190 residues, β helix | Receptor tyrosine kinases |
| C (furin-like, cysteine- rich) domain | ~120 residues, several small β hairpins, conformation stabilized by disulfide bonds | Receptor tyrosine kinases |
| Ig (immunoglobulin- like) | ~100 residues, antiparallel β sandwich, very similar to Fn3 | Receptor tyrosine kinases, many molecules in the immune system |

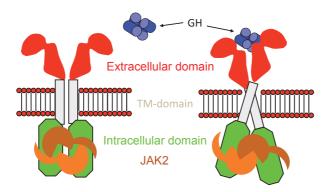


Fig. 14.2 ■ The binding of the growth hormone (GH) leads to a conformational change of its receptor. The tyrosine kinase JAK2 bound to the receptor dimer becomes activated and phosphorylates the cytoplasmic domains of the receptor.

monomer binds the ligand with high affinity. The receptor monomer-ligand complex presents a sufficiently large interacting surface consisting of elements of both the receptor and the hormone to which the second receptor molecule will bind after the conformational change.

The phosphorylase JAK2 binds to the intracellular part of the receptor close to the TM domain and is composed of several domains, one is a kinase domain and one a pseudokinase domain. The latter has a regulatory function. When bound to the receptor dimer without the bound hormone, the pseudokinase domain inhibits the kinase domain. When the hormone

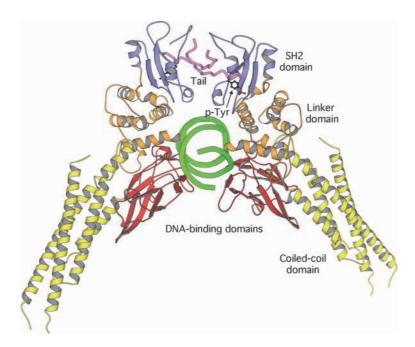


Fig. 14.3 ■ Intracellular activity: A complex between a dimer of STAT3 and a segment of dsDNA (green). In this view, the two-fold axis is close to vertical. The phosphorylated tyrosine (p-Tyr) mediating the dimerization is found in a C-terminal tail (lilac) that is bound in the SH2 domain of the other monomer (blue). The connection between the SH2 domain and the tail is disordered. The DNA-binding domain (red) has an immunoglobulin-type-fold similar to the p53 transcription factor (Section 10.2.6). It is preceded in the sequence by the coiled-coil domain (yellow) and followed by a linker domain (orange). The N-terminal domain is not included in this model (PDB: 1BG1).

binds, the intracellular parts of the receptor obtain the correct orientation to activate the bound JAK2 molecules. Here, the two kinase domains will be juxtaposed and transactivate each other to phosphorylate tyrosines of the cytoplasmic parts of the receptor.

The STAT molecules are a family of multidomain transcription factors that transmit the signal. STAT has an SH2 domain that binds specifically to phosphorylated tyrosine residues (see below). This domain binds to phosphorylated tyrosine residues of the receptor. A tyrosine residue close to the C-terminus of STAT also becomes phosphorylated by the JAK2 kinase. This causes the STAT molecules to dissociate from the receptor and dimerize through their SH2 domains. Subsequently, they are transferred to the nucleus where they bind to the DNA through their DNA-binding domains and thereby regulate the transcription of specific genes (Figure 14.3).

14.2.2 Control of Kinase Activity

Up to 30% of all proteins are phosphorylated at any given time. The human genome encodes in excess of 500 different protein kinases that take part in signaling pathways

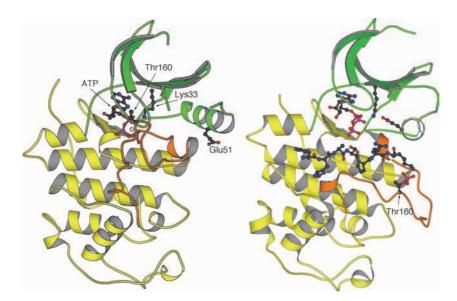


Fig. 14.4 ■ Inactive and active forms of cyclin-dependent kinase with the two domains of the enzyme in green and yellow. The activation segment is in orange. In the inactive form (*left*), it blocks the active site, preventing the binding of ATP. A conserved Glu51 points away from the active site. In the active form (*right*), with bound cyclin (not shown), the activation loop, where Thr160 is phosphorylated, has moved away. The C helix of the upper domain has rotated and moved towards the active site. Glu51 interacts with Lys33 and is close to the ATP molecule. A substrate peptide with its serine is close to the γ -phosphate of ATP is shown in atomic detail close to ATP (PDB: 1HCK and 1QMZ).

and are able to phosphorylate tyrosine or serine/threonine residues. This large family of protein kinases has two domains: the N-terminal domain that consists of an antiparallel five-stranded sheet and a helix, the C helix and the C-terminal domain containing seven helices (Figure 14.4). The active site is in a cleft between the domains.

The control of the activity of tyrosine kinases depends on a flexible loop called the activation segment. In inactive forms of the enzyme it is often partially disordered and blocks the active site. In active forms, it is ordered and allows the substrate to bind in the correct orientation. The conformational differences between active and inactive forms are illustrated for one kinase, cyclin-dependent kinase (Figure 14.4). In the inactive form, the C helix is far from the active site. In the active form (when a cyclin is bound to the kinase) the blocking activation segment moves out and helix C moves into the active site. In many kinases, activation is caused by phosphorylation of a residue in the activation loop, in this case a threonine. Examples of such activation are also discussed in Sections 14.2.3.1 and 14.3.1.

The removal of the phosphates is handled by protein phosphatases. In mammals, there are several hundred different variants, all with specific substrate proteins.

14.2.3 Control of Activity Using Binding Domains

Many proteins are modular i.e. they contain several domains. The function of multidomain proteins is to bring several activities to a desired place in the cell — to the targets where a range of enzymatic activities are needed, to the cell membrane or to other specific places (Table 14.3).

TABLE 14.3 Examples of Intracellular Binding Domains

| Domain Name | Structure | Binding Specificity | Examples of Proteins in Signal Pathways Containing the Domain |
|---------------------------------------|--|--|---|
| SH2 (Src homology domain 2) | ~100 residues, antiparallel β-sheet with a helix on each side (Figure 14.6) | Phosphotyrosine- containing peptides | Phospholipase C γ-isoform, tyrosine kinases (src), phosphatidylinositol-3- kinase, tyrosine phos- phatase, RasGAP, GRB2 |
| SH3 (Src homology domain 3) | ~50 residues, small β barrel (Figure 14.6) | Proline-rich peptides | Phospholipase C, tyrosine kinases (src), cytoskeletal proteins, RasGAP, GRB2 |
| PH (Pleckstrin homology) | ~120 residues, a β sandwich with a C-terminal helix | Membrane association and other specificities | Phospholipase C, protein kinases, GEF pro- teins (Sos), GAP proteins, cytoskeletal proteins |
| FERM | ~300 residues in three subdomains | Membrane association | Tyrosine kinases, protein phosphatases, cytoskeletal proteins (ezrin, moesin, radixin) |
| PTB (phosphotyrosine binding domain) | Similar to PH domain | Phosphotyrosine- containing peptides | Shc adaptor protein, IRS-1 insulin receptor substrate |
| C2 (protein kinase C domain) | ~120 residues, Ig-fold | Membrane association | Phospholipase C, protein kinase C, other kinases, RasGAP |
| C1 (protein kinase C domain) | ~50 residues, cysteine- rich domain binding two Zn ions | Diacyl glycerol, phorbol esters | Protein kinase C, other kinases |
| PDZ (DHR domains, GLGF repeats) | 80 residues, β barrel | Short peptides often ending with valine | Many proteins in signal systems, often present in tandem |
| WD40 | 7-bladed propeller | Various | β subunit of trimeric G-protein, Arp2/3 complex |
| EF hand | ~25 residues | Calcium ions | Phospholipase C with 4 EF hands |

14.2.3.1 SH2 and SH3 domains

SH2 and SH3 domains are found in many proteins. These domains are named after the non-receptor tyrosine kinases src (src homology). Src and other similar kinases are associated with the cell membrane and activated in a number of signaling pathways. They contain a kinase domain, one SH2 and one SH3 domain (Figure 14.5).

The SH2 domain binds only to segments in a protein with a phosphorylated tyrosine. An SH2 domain consists mainly of an antiparallel sheet with one helix on each side (Figure 14.6). The peptide-binding surface is formed by the edge of the sheet and the side of the two helices. There are two pockets, one for binding the phosphorylated tyrosine and one for binding of residue Y+3 (as in YEEI). Different SH2 domains have different specificity and the specificity is mainly due to the nature of this pocket.

SH3 domains are small, five-stranded antiparallel β barrels (Figure 14.6). The function of SH3 domains, like SH2 domains, is to bind specifically to other proteins. Segments of

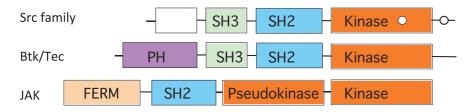


Fig. 14.5 ■ Domain structure of some non-receptor tyrosine kinases. Src has phosphorylation sites (marked with a little ball) within the kinase domain and in the C-terminal tail. For a description of PH, SH2, SH3 and FERM domains, see Table 14.3.

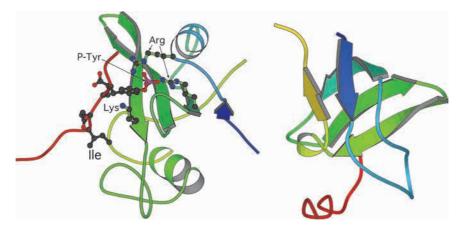


Fig. 14.6 ■ Left: The SH2 domain of src in complex with the peptide YEEI (red). Two arginines and a lysine residue are involved in binding the phosphate group of the phosphotyrosine (PDB: 1SPS.) Right: The peptide binding to an SH3 domain. This domain is from the Abl tyrosine kinase. The peptide bound (in red) has the sequence APTMPPPLPP and has a left-handed polyproline helix conformation (PDB: 1ABO).

these proteins bind in a groove between two loops. The sequence of peptides that bind is of the type RXLPPLPXX or XXXPPLPXR. These proline-rich peptides bind as a polyproline helix. The binding affinity is relatively low, and is in some cases increased by binding to multiple SH3 domains.

The Src tyrosine kinase and related kinases all have a C-terminal tail with one tyrosine phosphorylation site. When phosphorylated, this tyrosine binds in the phosphotyrosine pocket of the SH2 domain, but there is no binding in the second pocket, and the binding is therefore weak. This tail is important to regulate the activity. When the tyrosine is phosphorylated the protein is inactive. An oncogenic variant of src is found in Rous sarcoma virus, a virus causing tumors in chickens. This protein, v-src, lacks the C-terminal tail, and the kinase activity of the oncogene cannot therefore be regulated as the normal cellular src. This leads to uncontrolled growth and tumor formation.

The linker region between the SH2 domain and the kinase contains some proline residues and is bound to the SH3 domain. Inactive src shows the various domains in relatively close association (Figure 14.7).

The mechanism of activation is not completely obvious. From comparisons with other kinases, it appears that the kinase domains of src have a conformation where part of the active site (the helix of the first domain) has an orientation that prevents the catalytic

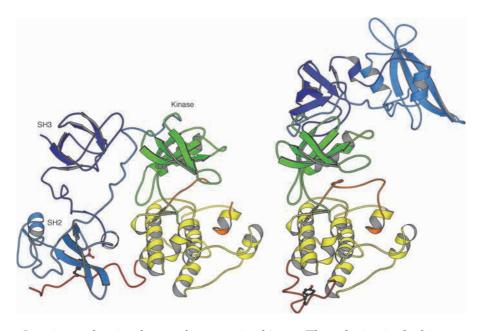


Fig. 14.7 ■ Inactive and active forms of src tyrosine kinase. The coloring in the kinase part is as in Figure 14.4. In the inactive form (*left*), the SH2 domain is bound to the phosphorylated tyrosine in the tail and the SH3 domain is bound to the linker. In the activated form (*right*), the tyrosine of the C-terminal tail is not phosphorylated, which allows the SH2 and SH3 domains (light and dark blue, respectively) to find a completely different orientation and be free to bind to other molecules (PDB: 1FMK and 1Y57).

activity. Binding of the tail and linker to the SH2 and SH3 domains stabilizes this inactive conformation; however, the SH2 and SH3 domains bind on the opposite side to the active site. Dephosphorylation of the tail or binding of peptides with high affinity to the SH2 and SH3 domains will disrupt the close association of the domains and allow the kinase part of the molecule to form an active enzyme. Another tyrosine in the activation segment is phosphorylated for full activity.

Receptor Tyrosine Kinase Pathways 14.3

A large group of signaling pathways uses receptors that have an intrinsic kinase activity with specificity for tyrosine residues. The human genome contains among 60 receptor tyrosine kinases (RTKs). This group includes the important insulin receptor and a number of receptors for growth factors. In these receptors, the receptor protein itself has tyrosine kinase activity, in contrast to the cytokine receptors, where the signal activates separate kinases.

Receptor tyrosine kinases consist of an extracellular receptor region, a single transmembrane helix, and a tyrosine kinase domain on the cytosolic side. The extracellular part of the receptor either contains a combination of a cysteine-rich domain and a leucine repeat domain or has various numbers of copies of Ig-type domains. Some of the receptors have unique domains in addition to the more common ones. Schematic drawings of some representative RTKs are shown in Figure 14.8.

Ligand binding to all receptors induces tyrosine kinase activity. In most RTKs, the binding on the extracellular side leads to dimerization of the receptor chains. However,

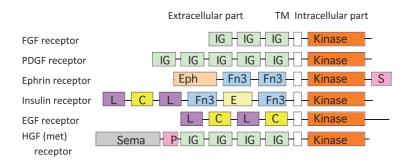


Fig. 14.8 ■ Domain organization of a number of receptor tyrosine kinases. They all have the intracellular kinase domain(s) and the transmembrane region (TM, small white rectangles) in common. Other modules are the immunoglobulin (IG) and fibronectin type III modules (Fn3), the leucine repeat domain (L), the furin or cysteine-rich domain (C), the SAM domain (S) and the ephrin ligand-binding domain (Eph). The insulin receptor has a couple of extra domains that have not been observed in other types of molecules and the hepatocyte growth factor or Met receptor has a large domain (Sema) with a seven-bladed propeller structure followed by a PSI domain (P).

The phosphorylated peptides of the receptor become targets for other proteins that are part of the signaling pathways. These proteins bind to the receptor tyrosine kinase, for example, through SH2 domains.

14.3.1 EGF-Receptor Pathway

The epidermal growth factor (EGF) uses the receptor tyrosine kinase pathway. The family includes four related receptors. They are of significant interest, since increased activity of the receptor can lead to certain forms of cancer. The activation of the receptor is connected to several pathways in the cell. Here, we will describe the activation of the MAP kinase pathway, but the JAK/STAT pathway (Section 14.2) and the protein kinase B pathway can also be activated by the EGF receptor (EGFR).

The receptor consists of two chains. Both homo- and hetero-dimers are observed. The monomer has four extracellular domains, a single transmembrane helix with the dimerization motif GxxxG (Section 4.6.3.2), and a cytoplasmic tyrosine kinase domain (Figure 14.8).

14.3.1.1 Extracellular part of EGFR

The extracellular domains are two leucine-repeat domains and two cysteine-rich domains (C domains) and about 40 kDa of carbohydrates, attached to 12 different sites. The structure is known for several forms of the extracellular part of the receptor (Figure 14.9). Domains I and III, also called L1 and L2 (for leucine-rich repeats), have a β helix fold. They form short rectangular rods, made up of repeating units of a parallel β -sheet. The C domains II and IV are rods of cysteine-rich repeats, where the conformation is dependent on the number of disulfide bonds. The extracellular part of the receptor appears to be a tandem duplication of the two domains.

In the monomeric form, the four domains have a bent conformation, but in some cases an extended conformation has been observed. In the dimeric form, the domains are rearranged to the extended conformation. Here, domains I and II are rotated by 130° compared to their positions in the bent conformation. Domain II, which in the monomeric form interacts with domain IV, now interacts with the other copy of domain II, partly using an extended loop (the dimerization arm) that is unique to this family of receptors. In contrast to other receptors, the ligand is not involved in the dimeric contact, but bound to domains I and III on the outside of the dimer. Ligand binding appears to stabilize the extended conformation of the receptor that forms the dimer.

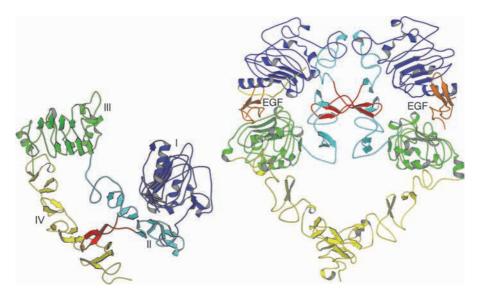


Fig. 14.9 ■ The structure of the extracellular part of the EGFR. *Left*: A monomer (PDB: 1NQL) and *right*: A dimer (PDB: 3NJP). The transmembrane helix is down in the figure. The loop in domain II that is important for dimerization is shown in red. Note that the ligand EGF is not directly involved in dimer formation.

14.3.1.2 The cytoplasmic part of EGFR

Binding of EGF to the receptor promotes dimerization and to autophosphorylation of EGFR at the cytoplasmic side. The phosphorylation of EGFR occurs in a C-terminal extension of more than 200 amino acid residues (Figure 14.8). The structure of the active form of the kinase domain was a surprise in being asymmetric. Here, one monomer is the enzymatically active subunit while the other monomer acts as an activator. A hydrophobic surface of the C-lobe of the activating kinase monomer interacts with a hydrophobic surface of the N-lobe of the active kinase monomer (Figure 14.10). The allosteric activation is here similar to the activation of the cyclin-dependent kinase (Section 14.2.2), where one monomer binds to the other in a way that leads to activation. Phosphorylation is not absolutely needed for activation, but the phosphorylated C-terminal extension can bind SH2 domains of proteins in the signaling pathways.

In other RTKs, the autophosphorylation often occurs in the activation loop that regulates the activity of the tyrosine kinase as described in Section 14.2.2. Here, the dimerization allows the monomers to phosphorylate the other subunit in *trans*. The flexibility of the activation loop allows it to temporarily move to unblock the active site, allowing ATP and the loop of the other subunit to enter the active site. Thus, the association of the subunits compensates for the low activity of the enzyme.

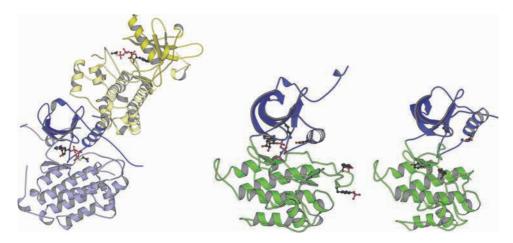


Fig. 14.10 ■ The tyrosine kinase domains of the EGFR. *Left*: The asymmetric dimer (PDB: 2GS6) where the active subunit is shown in blue and the activating subunit is shown in yellow. *Right*: The activated and inactive forms of the corresponding domain of the insulin receptor (PDB: 1IR3 and 1IRK).

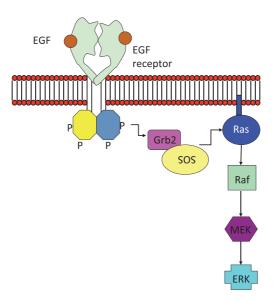


Fig. 14.11 ■ Overview of the activation of the Ras pathway by the EGFR. The kinases Raf, MEK and ERK are also called MapKKK, MapKK, and MapK, respectively. MapK (ERK) is transported into the nucleus, affecting transcription factors.

14.3.2 MAP Kinase Pathway

Many RTK signaling pathways, including the EGF pathway, use the G-protein Ras (Section 8.3.3), which activates a cascade of kinases called mitogen-activated protein kinases (MAPK, Figure 14.11). Many diseases of different kinds are due to abnormalities of the MAPK signaling system.

The first step in this pathway is the binding of the protein Grb2 to the autophosphorylated C-terminal part of the EGFR. Grb2 has two SH3 domains flanking an SH2 domain, which binds to phosphotyrosine residues in the receptor. Grb2 acts as an adapter and has no enzymatic function. Like other SH3 domains, the SH3 domains of Grb2 bind to proline-rich sequences. In this case, they bind to the protein Son of sevenless (Sos), which activates the membrane-bound G-protein Ras by acting as a G-nucleotide exchange factor (GEF) for Ras (Figure 14.11).

The membrane bound Ras with GTP attracts and activates Raf (also called MapKKK), a kinase with a Ras-binding domain. The activated Raf phosphorylates a serine residue in another kinase Mek (also called MapKK). This kinase in turn phosphorylates the kinase MapK (also called ERK, extracellular signal-regulated kinase). This kinase is transported into the nucleus to phosphorylate and activate specific transcription factors. In this way, EGF binding to the receptor leads to changes in the expression of specific genes. An activated protein generates a cascade of effects in the next step of the pathway until the protein is inactivated. The kinases are inactivated by specific phosphatases, while Ras is inactivated through GTP hydrolysis.

MapK and its regulatory proteins contain a docking site in the C-terminal lobe for other kinases called the kinase interaction motif (KIM), with a consensus sequence of 13–15 amino acids.

14.3.2.1 Activation of Ras through GDP release

All G-proteins have a common G domain (Section 8.3.3). The protein Ras is nothing but this G domain in its simplest form and is often referred to as the standard G-protein. Ras is associated with the plasma membrane through a fatty acid molecule added to a cysteine residue in the C-terminal part of the protein by a post-translational modification. The main GEF protein for Ras, Sos, is a relatively large protein with about 1400 amino acid residues. The region of Sos that is involved in the GEF activity for Ras, residues 560–1050, is helical and consists of two domains, one of which binds to the switch regions in Ras. The interactions are extensive (Figure 14.12). A helical hairpin of Sos is inserted between the switch I and II regions of Ras (Figure 14.13). Switch I gets displaced from the nucleotide binding site, and side chains from the Sos helical hairpin are inserted into the magnesium site and part of the GDP, leading to GDP release.

14.3.2.2 GTPase activity of G-proteins and the effect of GAP proteins

The GTPases normally function as molecular switches (Section 8.3.3). With GTP, they are in the ON state. The hydrolysis to GDP leads to the OFF state. The conformation of switch I and II differs considerably between the two states. The GTPase activity of G-proteins is normally low. A glutamine from switch II together with an arginine, the arginine finger, from

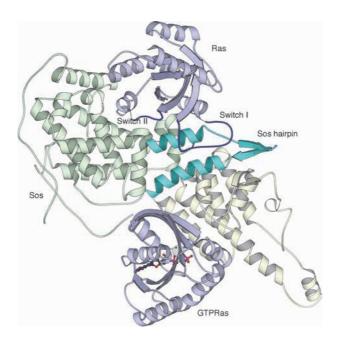


Fig. 14.12 ■ A complex between Sos and Ras. Sos is a GEF protein binding to its G-protein partner. Two of the domains of Sos are illustrated, the Ras-binding part, the Cdc25 homology domain (pale green) and the Rem (Ras exchanger motif) domain (pale yellow). A helical hairpin of the Cdc25 domain (turquoise) is inserted between switches I and II (dark blue) of Ras (pale blue), leading to release of GDP. A second Ras molecule in the GTP conformation binds to the Rem domain and has a regulatory role (PDB: 1XD2).

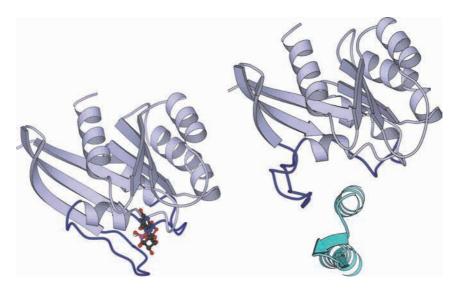


Fig. 14.13 ■ A comparison of Ras with a bound nucleotide (PDB: 121P) and the Ras-Sos complex (PDB: 1BKD), showing how the helical hairpin of Sos (*bottom right*) is inserted between switch I (to the *left* in both drawings) and switch II, and removes the nucleotide. The two switch regions are in dark blue.

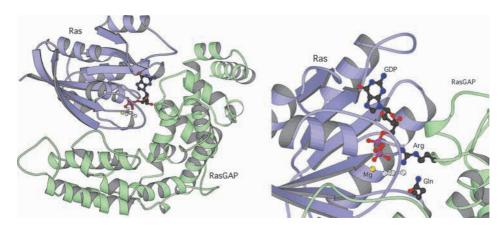


Fig. 14.14 ■ *Left*: A complex between RasGAP and Ras. Binding of RasGAP stabilizes the loops that are part of the active site. *Right*: Detail of the complex between Ras, RasGAP, GDP and AlF3, mimicking the transition state in GTP hydrolysis. RasGAP contributes the arginine residue involved in the GTP hydrolysis (the arginine finger; PDB: 1WQ1).

switch I are involved in the catalysis. In Ras and related proteins, the arginine is absent. In these cases, the GTPase-activating protein (GAP) contributes an arginine finger in *trans* at essentially the same position as the one found in other G proteins, increasing the hydrolysis rate of Ras by a factor of 1000.

RasGAP binds to the switches I and II in Ras (Figure 14.14). The binding of GAP places the active site residues in the right position for catalysis, which leads to the correct positioning and activation of the water molecule (Section 8.3.3.2). The negative entropic effect when these loops become less flexible is compensated by the binding energy of the GAP protein.

14.3.2.3 Ras and cancer

Certain mutations in Ras are associated with different types of cancer. These mutants leave Ras in a permanent ON state by preventing the GTP hydrolysis. Two mutants are especially common. If Gly12 in the P-loop is mutated to a valine, the activation of the GTPase activity through RasGAP is prevented. The second common mutation involves Gln61, a residue that is involved in correct positioning of the water molecule that participates in the catalysis.

14.4 G-protein Coupled Receptor, GPCR, Pathways

A large number of signaling pathways use a similar set of components to change the activity of proteins in the cell. One pathway uses a set of membrane proteins, formed by

14.4.1 Receptors

In humans, there are almost 800 different genes for GPCRs and this protein family is therefore the largest family in the genome. Well-known examples of such receptors are the adrenergic receptors, rhodopsin and various olfactory receptors (Table 14.4). Based on similarities in their amino acid sequences, the GPCRs can be divided into six families (Table 14.5). Most of the receptors belong to the rhodopsin-like GPCRs. Examples of other families found in humans are the secretin-like GPCR family (also called glucagon receptor-like) and the metabotropic glutamate receptor family. Most notably, these families differ in the size of the extracellular N-terminal parts that precede the transmembrane region. The rhodopsin-like receptors have a short N-terminal segment, while the

TABLE 14.4 Examples of the Main Components of Some G-Protein Coupled Receptor Pathways

| Receptor (GEF) | G-Protein | Effector | Class |
|---------------------------|--------------------------------|---------------------------------|-------|
| Rhodopsin | Gt (Transducin) | cGMP phosphodiesterase | A |
| β2-adrenergic receptor | Gs_{α} | Adenylate cyclase (stimulating) | A |
| CXCR4 receptor | Gi_{lpha} | Adenylate cyclase (inhibiting) | A |
| Acetylcholine M1 receptor | Gq_{lpha} | Phospholipase C | A |
| Odorant receptors | $\operatorname{Golf}_{\alpha}$ | Adenylate cyclase (stimulating) | A |
| Glucagon receptor | Gai | Adenylate cyclase | В |

TABLE 14.5 The Classes of GPCR Receptors

| Class | Name |
|-------|-----------------------------------|
| A | Rhodopsin-like |
| В | Secretin receptor family |
| C | Metabotropic glutamate/pheromone |
| D | Fungal mating pheromone receptors |
| E | Cyclic AMP receptors |
| F | Frizzled/Smoothened |

glutamate receptors have a large N-terminal region of upto about 600 residues that forms the extracellular surface of the receptor.

The GPCRs convert ligand binding at the outer surface of the cell membrane into conformational changes at the inside of the membrane. Rhodopsin has served as the prototype for this group of receptors, although rhodopsin is different from the other receptors in that it does not bind a soluble ligand to initiate the signaling pathway.

In general, it is difficult to crystallize integral membrane proteins and the GPCRs have been exceptionally difficult. The first GPCR structure determined was that of rhodopsin (Figure 14.15). A related protein with the same fold that has been studied since the 1970s is bacteriorhodopsin. Several other structures of GPCRs have subsequently been determined. All these GPCRs have a similar fold with seven transmembrane helices. The extracellular part consists of the N-terminus and three loops, and the intracellular surface is formed by the C-terminus and three loops. A comparison of the known structures shows that their extracellular parts differ considerably, reflecting that they bind ligands of very

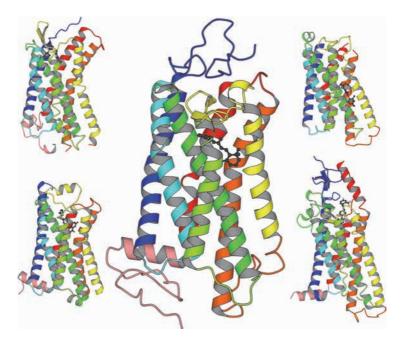


Fig. 14.15 ■ GPCR molecules are different in their extracellular (*up*) and intracellular parts (*down*). Some GPCRs were modified in the intracellular loops by an inserted protein domain to simplify crystallization. Such modifications are not shown in the drawings. Center: The rhodopsin structure (PDB: 1U19). Top left: CXCR4 receptor (PDB: 3ODU), bottom left: β2 adrenergic receptor (PDB: 3P0G), top right: corticotropin-releasing factor receptor with a truncated C-terminus (PDB: 4K5Y) and bottom right: smoothened receptor (PDB: 4JKV). The seven transmembrane helices are colored from blue to red. The C-terminus, which in most cases is a helix, is pink. Each structure shows a bound ligand.

different character (Figure 14.15). The intracellular part interacts with trimeric G-proteins, which are all similar, and this surface differs less between the GPCRs.

14.4.2 Trimeric G-Proteins: Switching

The G-proteins are molecular switches. The trimeric G-proteins in the GPCR signaling pathways work as amplifiers: when activated, by binding to the activated receptor, they can activate many effectors before they are inactivated. They consist of three chains: α , β and γ . In contrast to the large number of GPCRs, there are only about 20 genes for α chains and even fewer for β and γ chains. Each trimeric G-protein thus binds to many different GPCRs.

14.4.2.1 Trimeric G-proteins, structure

The structure of a trimeric G-protein is shown in Figure 14.16. The complex is associated with the membrane through lipid molecules that are covalently associated with the N-terminus of the α and the C-terminus of the γ subunits. Since both the activating receptors and the effector molecules that are influenced by the G-proteins are membrane-bound, this increases the efficiency of the G-proteins.

 $G\beta$ is a seven-blade propeller protein, where each blade is an antiparallel four-stranded β sheet (Figure 14.17). An N-terminal helix is followed by the seven-fold repeating motif of 36–40 amino acids that ends with relatively conserved tryptophan and aspartic acid residues (WD). The first β -strand is the outermost strand in the seventh and

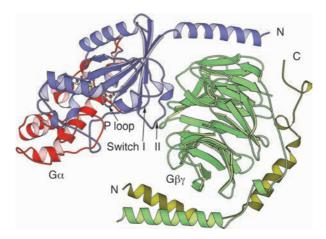


Fig. 14.16 • A Gαβγ complex of transducin (PDB: 1GOT). The α chain has a G domain (blue) with an inserted helical domain (red). Gβγ (Gβ is green and Gγ is olive) binds to the switch regions of the G domain.

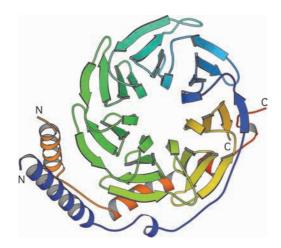


Fig. 14.17 • The $G_{\beta\gamma}$ complex from transducin. G_{γ} is shown in orange and red (PDB: 1GOT).

last blade. The conserved residues are important for the stability of the structure. G γ is a short segment that is probably disordered when isolated, but forms two helices that are an integral part of the protein when in contact with G β .

The binding of $G\alpha$ subunit to the $G\beta\gamma$ subunits involves the switch regions and locks them in the GDP conformation (Figure 14.16). The $\beta\gamma$ subunits stabilize the inactive GDP form of the α subunits. A disordered region at the N-terminus of α is also ordered, forming a α -helix in contact with $\beta\gamma$.

14.4.2.2 Trimeric G-proteins, mechanisms

The inactive GDP form of trimeric G-proteins binds to the GPCR receptor. The receptor has a resting state, but changes its conformation when a ligand (e.g. adrenaline) binds to the receptor, or when light photons hit rhodopsin. The activated receptor then acts as a GEF protein, inducing GDP to be released from the inactive G-protein and GTP to bind and activate the G-protein.

When GTP is bound, the activated G-protein dissociates into two parts, G α and G $\beta\gamma$ with low affinity for the receptor. The subunits dissociate from the receptor and associate with effector proteins to activate or inhibit them. The G-protein is active until the GTP is hydrolyzed to GDP. Thus, the duration of the signal depends on the GTPase activity of the G-protein. When GTP is hydrolyzed to GDP, G α and G $\beta\gamma$ again associate with each other and the reformed trimer can bind to the receptor to allow another cycle of activation.

The $G\alpha$ subunit of a trimeric G-protein has two domains, a GTPase domain and a helical domain (Figure 14.18). The GTPase domain is a classical G-domain similar to Ras. The helical domain is a long insert into switch I. This domain appears to function as a lid covering the bound nucleotide.

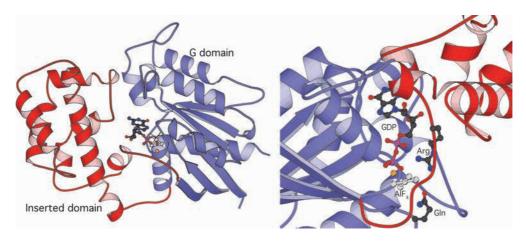


Fig. 14.18 ■ *Left*: The α subunit of Gs showing the G domain of transducin and its inserted helical domain (PDB: 1TAD). Right: The helical domain acts as an internal GAP protein, contributing an arginine (the Arg-finger) to the active site. The complex with GDP and an aluminum fluoride ion, AlF₄, represents a transition state analog in the hydrolysis of GTP. The Gln residue that participates in placing the hydrolytic water molecule is also shown.

The helical insertion domain (Figure 14.18) of the trimeric G-proteins contributes the arginine (Arg finger) involved in the GTPase activity. This helical domain can thus be regarded as an internal GAP domain. It can indeed act in that way even if it is separated from the G-protein by genetic engineering and supplied as a separate protein.

The G domain has three regions that change conformation in relation to whether GTP or GDP is bound. Two of these correspond to switches I and II in Ras. The conformational differences caused by the presence or absence of the γ phosphate are considerable. In the GDP form, a serine from the P-loop and an oxygen atom from the β phosphate coordinate the Mg²⁺ ion. In the GTP form, a threonine from switch I and an oxygen atom from the γ phosphate also become ligands. The presence of the γ phosphate thus moves the switch I region towards the magnesium ion and the phosphate (Figure 14.19).

The conformational changes in switch II are due to direct interactions with the γ phosphate. A main chain nitrogen from switch II forms a hydrogen bond to the phosphate, changing the orientation of the β strand and helix connected by this loop.

14.4.3 Visual System

In visual transduction, the receptors are proteins called opsins. A chromophore, retinal, is bound to the protein. In rod cells, the complex between the apoprotein opsin and retinal is called rhodopsin. The rhodopsin molecule is sensitive to photons (Figure 14.20). When it is activated by light, the retinal molecule changes isomerization from 11-cis-retinal to all-trans-retinal. This leads to conformational changes in the rhodopsin molecule,

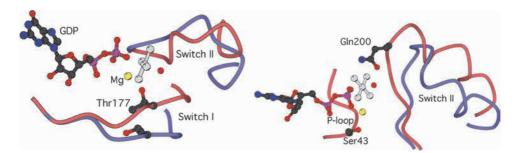


Fig. 14.19 \blacksquare Conformational changes at the magnesium ion and the γ phosphate. *Left*: Thr177 binds to the magnesium ion and the γ "phosphate" in the GTP conformation (red, PDB: 1TAD). As in other structural studies of GTPases, the GTP conformation is obtained through binding of molecules mimicking GTP or the transition state, in this case GDP and AlF $_4$. In blue is the conformation of the switch I and II loops in the GDP form (PDB: 1TAG). *Right*: Switch II changes conformation due to new contacts between the oxygen atoms of the "third phosphate" when a GTP analog is bound. This leads to a shift of the switch II helix. Ser43 is part of the conserved GXXXXGKS motif in P loop-containing proteins.

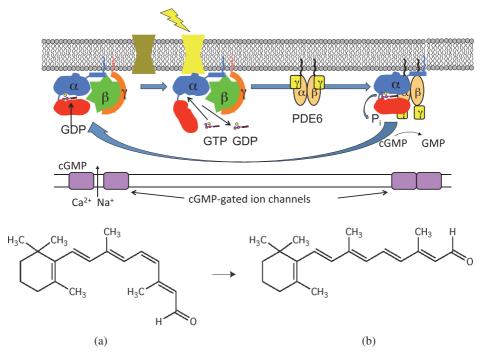


Fig. 14.20 Top: A view of the signaling pathway used in rod cells in the eye. Before the signal, cGMP keeps the cGMP-gated ion channel open. When rhodopsin is activated by photons (flash), the trimeric G-protein transducin is activated by rhodopsin (yellow), exchanging its GDP for GTP. The Gα subunit is released, to interact with the cyclic GMP phosphodiesterase (PDE6) and displace the inhibitory γ subunit. This leads to hydrolysis of cGMP and closing of ion channels. *Bottom*: The retinal molecule and its light-induced conformational change. (a) 11-*cis* retinal and (b) The excited all-*trans*-retinal.

the G-nucleotide exchange protein (GEF) for transducin inducing release of GDP and binding of GTP. Activated rhodopsin generates an important amplification step by activating hundreds of transducin molecules. The activated transducin in turn activates a cyclic GMP phosphodiesterase by removing its inhibitory γ subunit. This enzyme degrades cyclic GMP, which closes cGMP-gated Na⁺ channels. In this process, transducin immediately hydrolyses its GTP molecule and has to be reactivated by rhodopsin. Thus, light photons hitting rhodopsin molecules in the retina lead to a change in membrane potential and the level of secreted neurotransmitters causing a neural impulse, enabling us to see.

14.4.3.1 Rhodopsin structure

Rhodopsin is exceptional in the way that it does not bind an external ligand; instead the signal is a conformational change in a covalently bound retinal molecule. The retinal molecule is a hydrocarbon molecule consisting of a tail and a ring structure (Figure 14.21). The tail is covalently linked to a lysine residue in the protein. The light activation switches the retinal from the 11-cis to the all-trans form. The energy barrier in this transition is high, thus preventing basal signaling without any light input. This change influences the arrangement of the helices and the loops on the inside of the cell in a way that affects the binding of the G-protein transducin (see Section 14.4.4). The effects of the conformational changes in the light-sensitive retinal molecule correspond to the effects that in other receptors are caused by binding of a ligand — the photon is the "ligand" that "binds" to the covalently attached retinal.

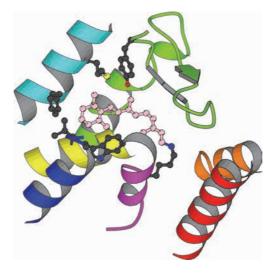


Fig. 14.21 ■ Retinal (pink) bound to rhodopsin with the ring structure bound in a hydrophobic pocket. The tail is covalently bound to a lysine side chain (PDB: 1F88).

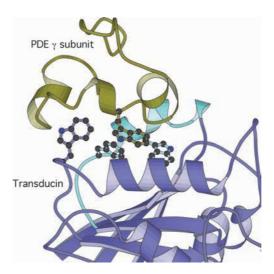


Fig. 14.22 ■ Transducin binding to the γ subunit of cGMP phosphodiesterase (PDB: 1FQJ). The switch II loop and helix is turquoise. The tryptophan of the phosphodiesterase interacts with conserved residues at the switch II of transducin α subunit.

14.4.3.2 *Effector*

The effector molecule regulated by transducin is cyclic GMP phosphodiesterase (PDE6), which is associated with membranes in rod cells. PDE6 has three subunits, α , β and γ . The γ subunit acts as an inhibitor of the enzyme. The active form of the α subunit of transducin binds to the C-terminal part of the phosphodiesterase γ subunit, preventing it from inhibiting the enzyme. The enzyme becomes active and cGMP will be degraded (Figure 14.22). The rhodopsin signal transduction is very rapid, in the order of milliseconds.

In the complex between the PDE6 γ subunit and transducin, an aromatic residue in the effector molecule is inserted between the switch II helix and the α 3 helix of the G-protein, contacting conserved residues. This interaction depends on the nucleotide status of the G-protein, since the switch II helix has a different orientation in the GDP conformation. Thus, when the GTP of transducin is hydrolyzed, the PDE6 γ subunit will dissociate and again inhibit its enzyme.

14.4.4 **\beta2-Adrenergic Receptor**

The $\beta 2$ -adrenergic receptor has been a model system for studying GPCRs. It is one of nine genes of related neurotransmitters. The neurotransmitter epinephrine or adrenalin regulates the heart rate and blood pressure. Propanolol is a β -blocker and one of the drugs used in the treatment of conditions relating to heart activity and blood pressure.

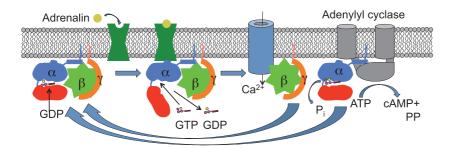


Fig. 14.23 ■ The β 2-adrenergic receptor (β_2AR) and its trimeric G-protein Gs. The process starts on the left where the inactive Gs-protein is attached to the membrane. When an agonist (adrenalin) binds to the receptor, it changes conformation to enable Gs to bind and replace its GDP for GTP. This makes the α subunit dissociate from the β and γ subunits. The α subunit activates the adenylyl cyclase, which produces cyclic AMP from ATP. The β and γ subunits activate a calcium channel.

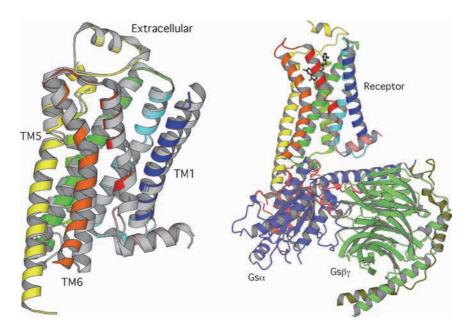


Fig. 14.24 • Left: The conformational changes of $β_2AR$ due to the binding of a high affinity agonist. The active form (in color) is from the complex with the trimeric G-protein (PDB: 3SN6). The inactive form (PDB: 2RH1) is shown in grey. TM6 is moved to accommodate for the binding of the G-protein. *Right*: The complex between the β2 adrenergic receptor and the Gs-protein (PDB: 3SN6). The fifth transmembrane helix of the receptor (yellow) is extended and interacts with the α subunit of the G-protein.

The activation of GPCR pathways has been possible to study due to the crystal structure of a complex between the β 2-adrenergic receptor (β_2 AR) and its trimeric G-protein (Gs, Figures 14.23 and 14.24). The structure of the complex was obtained after significant efforts and showed the conformational changes the GPCR undergoes due to the binding of the agonist. This remains the only structure of a complex between a GPCR and its G-protein.

The β_2 AR, like many GPCRs, is a pharmaceutical target for the treatment of various diseases. The activation by its natural agonist adrenaline and the signal transmission for β_2 AR is much less efficient than for rhodopsin. Adrenaline binds to β_2 AR with relatively low affinity (Ki around 1 µM). Due to the low affinity and rapid dissociation it is not certain that each agonist binding will generate a signal.

The binding of the agonist to β_2 AR leads to a drastic conformational change, related to the dynamic character of the protein. The fifth TM helix (TM5) is extended by a few helical turns and TM6 moves outwards by 14 Å (Figure 14.24). This generates a new pocket into which the trimeric G-protein can bind.

There is no crystal structure of this trimeric G-protein alone, but a comparison with known trimeric G-proteins shows that the C-terminal helix of the G-protein moves to interact strongly with the receptor. The other end of this helix reaches the nucleotidebinding site and conformational changes caused by ligand (agonist) binding to the receptor may lead to GDP release in the G-protein. The most striking conformational difference is that the helical domain in Gsα has moved to a new position in the nucleotide-free trimeric G-protein (Figure 14.25). This removes the lid from the bound nucleotide and indicates a considerable flexibility of the molecule.

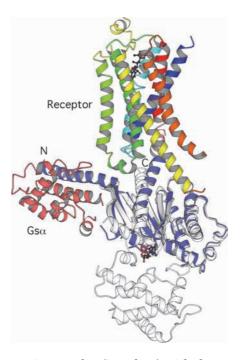


Fig. 14.25 \blacksquare The receptor-G-protein complex (in colors) with the structure of Gs α in the GTP conformation (PDB:1AZT, grey) superimposed. The position of the GTP analog in this structure is shown. The helical domain of Gsα in the receptor complex (red) has a completely different conformation than in the G-protein itself. This leaves the nucleotide site wide open for exchange. The C-terminal helix of the G-protein has changed conformation to interact with the receptor.

The main second messenger in the G-protein-coupled receptor pathways is cyclic AMP (cAMP). This molecule is produced by the enzyme adenylyl cyclase. There are nine different human isoforms of adenylyl cyclase. The enzyme forms cAMP from ATP and is regulated by a number of signaling pathways. The best-characterized activation of an effector in a GPCR pathway is that of adenylyl cyclase by the G-protein $G_{s\alpha}$. This pathway starts with the activation of a number of receptors, for example, the β -adrenergic receptor. These receptors all act on the trimeric G-protein Gs, where the "s" stands for stimulatory.

The low basal activity of adenylyl cyclase is enhanced by $G_{s\alpha}$, the α subunit of the stimulatory Gs complex. Other activators work on different isoforms of adenylyl cyclases, and in this way different cell types can have a specific response due to the specific isoforms expressed in that particular cell type. For example, $G_{i\alpha}$ inhibits some adenylyl cyclases, in spite of $G_{s\alpha}$ and $G_{i\alpha}$ being very similar molecules.

Adenylyl cyclase is a single polypeptide consisting of a short N-terminal cytosolic segment followed by two similar modules. Each module has a membrane domain consisting of six transmembrane helices and a cytoplasmic catalytic region. The cytoplasmic regions are called C_1 and C_2 , each containing two domains. One domain in each segment, C_{1a} and C_{2a} , is homologous and catalytically active and the conformations of the domains are very similar. The two domains form a heterodimer. Constructions including only the C_{1a} and C_{2a} domains are catalytically active. The domains have a central

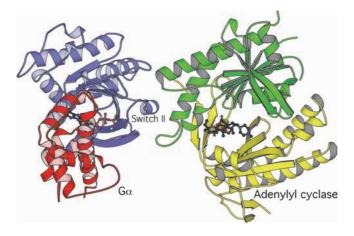


Fig. 14.26 ■ The complex between $G_{s\alpha}$ and the catalytic part of adenylyl cyclase. The two subunits of the enzyme are C1a (yellow) and C2a (green). The binding surface on adenylyl cyclase is designed so that only the GTP conformation of Gsα will bind and activate the enzyme. A GTP analog is bound to the Gsα subunit. Forskolin, a complex molecule that stimulates the adenylyl cyclase activity is also bound to the enzyme (PDB: 1AZS).

four-stranded antiparallel sheet with helices on both sides. An arm subdomain is formed by extensions of two of the strands, and there is also a C-terminal helical subdomain. The topology of the sheet is the same as in DNA/RNA polymerases, where ATP is bound in the same place and two "conserved" Asp residues bind the two Mg^{2+} ions, which in turn bind to the phosphates. The active site is located at the interface between the domains, with parts of both domains involved in catalysis. The crystal structure of a complex between C_{1a} and C_{2a} domains, $Gs\alpha$ and another activator, forskolin, is known (Figure 14.26). The helix of the switch II region of $Gs\alpha$ binds in a groove of the cyclase. In the GDP conformation, this helix is at a position that does not allow binding to the enzyme. Only the activated GTP form of the G-protein will therefore be able to activate the cyclase.

For Further Reading

Original Articles

- Boriack-Sjodin P, Margarit SM, Bar-Sagi D, Kuriyan J. (1998) The structural basis of the activation of Ras by Sos. *Nature* **394**: 337–343.
- Brooks AJ, Dai W, O'Mara ML, *et al.* (2014) Mechanism of activation or the protein kinase JAK2 by the growth hormone receptor. Science **344**: 1249783.
- de Vos AM, Ultsch M, Kossiakoff AA. (1992) Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science* **255**: 306–313.
- Lambright DG, Noel JP, Hamm HE, Sigler PB. (1994) Structural determinants for activation of the α-subunit of a heterotrimeric G protein. *Nature* **369**: 621–628.
- Lambright DG, Sondek J, Bohm A, et al. (1996) The 2.0 Å crystal structure of a heterotrimeric G protein. *Nature* **379**: 311–319.
- Lu C, Mi LZ, Grey MJ, et al. (2010) Structural evidence for loose linkage between ligand binding and kinase activation in the epidermal growth factor receptor. *Mol Cell Biol* **30**: 5432–5443.
- Ogiso H, Ishitani R, Nureki O, *et al.* (2002) Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell* **110**: 775–787.
- Palczewski K, Kumasaka T, Hori T, *et al.* (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**: 739–745.
- Rasmussen SG, DeVree BT, Zou Y, et al. (2011) Crystal structure of the beta2 adrenergic receptor-Gs protein complex. *Nature* **477**: 549–555.
- Tesmer JJ, Sunahara RK, Gilman AG, Sprang SR. (1997) Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gsα.GTPγS. *Science* **278**: 1907–1916.
- Xu W, Harrison SC, Eck MJ. (1997) Three-dimensional structure of the tyrosine kinase c-Src. *Nature* **385**: 595–602.

Reviews

Hubbard SR, Miller WT. (2007) Receptor tyrosine kinases: Mechanisms of activation and signaling. Curr Opin Cell Biol 19: 117-123.

Kovacs E, Zorn JA, Huang Y, et al. (2015) A structural perspective on the regulation of the epidermal growth factor receptor. Ann Rev Biochem 84: 739-764.

Oldham WM, Hamm HE. (2007) How do receptors activate G proteins? Adv. Prot Chem 74: 67–93. Palczewski K. (2006) G protein-coupled receptor rhodopsin. Ann Rev Biochem, 75: 743–767.

Scheffzek K, Ahmadian MR. (2005) GTPase-activating proteins: Structural and functional insights 18 years after discovery. Cell Mol Life Sci 62: 3014–3038.

Wells JA, Kossiakoff A. (2014) New tricks for an old dimer. Science 344: 703-704.

Cell Motility and Transport

In eukaryotic cells, there are many systems for maintaining the correct shape or motility of the cell. For these purposes, the cell has a system of fibers, the cytoskeleton. There are also systems for transporting material from the endoplasmic reticulum or the cell membrane to other parts of the cell. Vesicles that are transported along fibers in the cytoskeleton handle this traffic of proteins and other molecules. In addition, in muscle cells of many animals, there are fiber-forming molecules that allow the organism to make controlled movements. Some of these systems correspond to similar, simpler versions found in bacteria.

15.1 Actin Microfilaments

15.1.1 Actin: A Protein that can Form a Dynamic Fiber

Actin monomers have a molecular mass of about 42 kDa. In vertebrates, there are three isoforms of actin: α , β and γ . α -actin is a main component of muscle fibers, where it forms the thin filament. The β and γ forms are the most abundant proteins in most non-muscle eukaryotic cells, where they form a main component of the cytoskeleton. The actin microfilaments of the cytoskeleton are essential for the motility of cells as well as internal cell motility and cell division. Obviously, the microfilaments need to remodel themselves in a highly dynamic fashion.

Actin exists in two forms, G- and F-actins. F-actin is a fibrous form of actin, where actin molecules form long helices. G-actin (for globular) is the monomeric form. The two forms exist in equilibrium with each other: G-actin spontaneously forms F-actin above a certain concentration, and at low concentration the actin fiber dissolves. F-actin is polar: the two free ends of the polymer are different. Growth of the F-actin fiber occurs mainly in one

Actin binds ATP, which is important for the growth of actin fibers. Only actin with bound ATP is added at the "+" end of the fiber. The growing end of the fiber has a number of ATP-containing actin monomers before ATP is hydrolyzed to ADP. The phosphate produced in the hydrolysis is released very slowly, but actin monomers closer to the pointed end have lost the phosphate. This means that ATP-actin is added in the "+" end and ADP-actin is lost in the "-" end. The hydrolysis of ATP is a timing device that controls the degradation of the fiber.

Actin polymerization is important for many cellular properties, for example, shape and motility. Therefore, the growth of F-actin fibers in the cell is controlled by a number of proteins. Some block the polymerization or de-polymerization (capping proteins) and others are able to cut the actin fiber into shorter fragments (severing proteins). Many proteins are involved in cross-linking actin fibers in muscle and in the cytoskeleton. Some of these proteins are involved in forming bundles of actin fibers, others form crosslinks in a three-dimensional network of fibers.

In humans, there are six different actin genes, which are expressed differently depending on the cell type. The alpha chain is found in muscle. The beta and gamma actin molecules coexist in most cell types as components of the cytoskeleton and are involved in internal cell motility.

When actin molecules from various species are compared, only very few changes are seen in the sequences. Actin is one of the most conserved of all proteins. In actin and other conserved proteins, the protein surface is important because of all the specific interactions that occur at the surfaces with other proteins. Therefore, the amino acid sequence has changed only very little during evolution.

15.1.1.1 The actin monomer

The actin monomer is a relatively compact protein. Actin has two halves, each with two domains (Figure 15.1). Domain 2 is an insertion in domain 1, and domain 4 in domain 3. The C-terminal 40 residues after domain 3 in the sequence form a few helices that are part of domain 1. The active site where ATP is bound is in a deep crevice between the two halves of the molecule. Domains 1 and 3 have the same topology, a parallel sheet with helices on both sides. The same fold is also found in some seemingly unrelated proteins, for example, the chaperone hsp70 (see Section 12.1.2.3), hexokinase and some other kinases. Although the functions of these proteins are very different, all have an ATP-dependent activity in common.

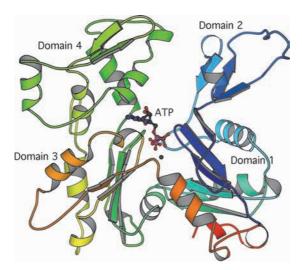


Fig. 15.1 ■ Schematic drawing of an actin monomer. The "–" end is at the top. The coloring is from N-terminus (blue) to C-terminus (red) (PDB: 1ATN). The long loop (D-loop) in domain 2 is often disordered in crystal structures of actin and actin complexes.

The hexokinase molecule is a well-known example where binding of the substrate glucose induces domain rotations. This type of fold may be suitable for domain rotations. Such flexibility could be important for the control of fiber growth, ATP hydrolysis, release of ADP and binding of ATP, and other mechanisms in the formation and degradation of actin networks. The structure of many forms of actin is known, but the domain rotations observed are not as large as in hexokinase. It is difficult to crystallize a flexible molecule with a strong tendency to form fibers, and it may therefore be difficult to decide to what extent a particular conformation is caused by the procedures used for crystallization. For example, structures of actin with ATP or ADP bound are similar, in spite of their different affinities to fibrous actin.

15.1.1.2 Structure of the actin fiber

Actin fibers form a superhelix of two actin helices, which have been studied with fiber diffraction and high resolution cryo-electron microscopy (Figure 15.2). If the structure is considered as a single helix, the rotation from one monomer to the next is 166.4° and the rise per subunit 27.5 Å. Domain 2 is the most exposed part of the actin molecule. The monomers form contacts both along the helix and between the two "strands" of the superhelix. The conformation of F-actin is different from G-actin: domain rotations lead to a more flat monomer in the fiber.

The low ATPase activity of monomeric actin is stimulated in the fiber. The mechanism for this stimulation in the fiber is not known but it may be similar to the effect of some

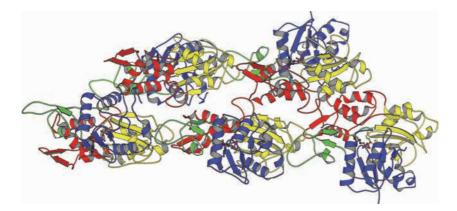


Fig. 15.2 ■ A model of the actin fiber. Five actin monomers are shown, three in the lower strand and two in the upper strand of the actin double helix. The "-" end (up in Figure 15.1) is to the left. Domain 1, blue, domain 2, green, domain 3, yellow and domain 4, red. Domains 1 and 2 are more exposed than domains 3 and 4 (PDB: 3J8A).

GAP proteins on G-proteins, where the enzymatically active conformation is stabilized by the contacts with other molecules.

15.1.2 Regulation of Actin Polymerization

15.1.2.1 Proteins binding to monomeric actin

There are a large number of proteins that bind to actin. At least 100 proteins control the polymerization of actin (Table 15.1). These proteins either bind to monomeric actin, prevent growth (capping) of actin fibers or sever the fibers.

15.1.2.2 *Profilin*

Together with thymosin β4, the protein profilin keeps actin in monomeric form at a concentration that is orders of magnitude higher than for barbed end polymerization. Profilin recharges ADP-actin with ATP. Profilin binding results in a moderate rotation of the two halves of the actin monomer, opening the cleft to allow ADP to leave and ATP to bind. The crystal structure of the complex of actin with profilin shows that it binds to the "+" end. Thus, actin-profilin complexes will only add to the barbed ("+") end of actin fibers (Figure 15.3).

15.1.2.3 Thymosin $\beta 4$ and Spire

The 43 aa peptide thymosin β 4 binds actin in cells. It is present in large amounts and is responsible for maintaining a high amount of monomeric actin. The protein does not have

TABLE 15.1 Examples of Proteins Involved in the Control of Actin Polymerization

| Name | Function | Conformation of Actin-Binding Domain |
|---------------------|---|--------------------------------------|
| Thymosin β4 | Storing monomeric actin. | Short peptide with the WH2 motif. |
| Spire | | Large protein with four WH2 motifs. |
| Formin (Cappuccino) | | Contains FH1 and FH2 domains. |
| Profilin | Binds to "+" end of monomeric actin, nucleotide exchange. | Profilin domain. |
| Gelsolin | Capping at "+" end of fiber, severing. | Gelsolin domain. |
| Arp2/3 complex | Binding to "-" end and nucleating new "+" ends at the side of actin fibers. | Actin-like. |



Fig. 15.3 ■ Schematic drawing of the profilin-actin complex (PDB: 1HLU). The "+" side of actin is blocked by profilin (red).

a compact globular structure in solution. When binding to actin, an N-terminal helix binds at the "+" end, while another helix binds at the opposite end of the protein. An extended segment connects these helices. This small protein is thus able to cap both ends of an actin monomer and prevent polymerization.

Many proteins involved in the regulation of actin assembly have a small module (17–27 aa) called WH2 (Wiskott-Aldrich syndrome protein or WASP homology 2).

This module corresponds to the N-terminal helix of thymosin $\beta 4$ that binds to the cleft of the barked end between domains 1 and 3 of actin (Figure 15.4). The C-terminal part, after an extended region, has four well-conserved amino acid residues, L++T/V (+ stands for K or R), binding in a hydrophobic groove. The basic residues interact with neighboring acidic residues.

Essentially, all proteins involved in actin filament nucleation or elongation use the WH2 domain. In filament-nucleating proteins it often occurs as tandem repeats. Spire is an example with four WH2 repeats (Figure 15.5). The WH2 repeats bind and organize actin monomers in a way similar to their orientation in the polymer. In addition, Spire has a C-terminal domain called KIND (kinase non-catalytic C-lobe domain) and a domain called FYVE, which is a zinc finger domain with eight cysteines binding two zinc ions and binding at the cytoplasmic membrane.

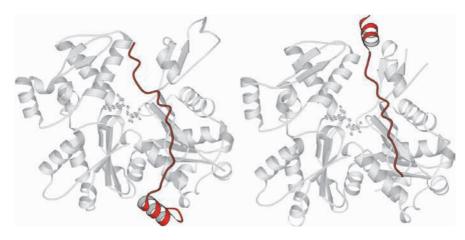


Fig. 15.4 ■ Fragments of thymosin $\beta 4$ (red) blocking the polymerization of monomeric actin. *Left*: The binding of a WH2 module to actin. The WH2 motif is part of the WASP-homology 2 protein (PDB: 2A41). An N-terminal helix binds at the target-binding groove between domains 1 and 3 in actin and an extended portion binds along the actin molecule. *Right*: Binding of the C-terminal of thymosin $\beta 4$ to the "–" end of actin (PDB: 1T44). Together, these two fragments represent the binding of thymosin $\beta 4$.

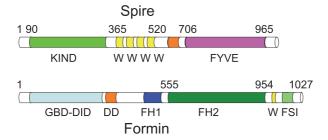


Fig. 15.5 ■ The domain arrangement of Spire and formin. W stands for WH2 domains.

15.1.2.4 Formin

Several proteins are engaged in the nucleation of actin fibers and a collective name for one family is formin. In mammals there are 15 different formins, which are large multidomain dimers with two distinctive domains, formin homologies 1 and 2 (FH1 and FH2, Figure 15.5). FH1 is a proline-rich region that can interact with profilin to recruit profilinactin complexes for the elongation of filaments. The FH2 domain occurs after the FH1 domain, but just before the C-terminus of formin. FH2 is responsible for the formation of the double filament arrangement of actin.

The FH2 domain of formin has five subdomains or regions. They have the names lasso, linker, knob, coiled-coil (triple-helical) and post from amino to carboxyl end. One crystal structure shows a dimer of FH2 forming a ring around two actin monomers, all related by a two-fold axis. The actin molecules do not contact each other but bind to FH2 (Figure 15.6). The lasso is an unusual arrangement for the two FH2 monomers to be connected. It extends from the knob through the linker to the lasso region that reaches the post of the other monomer and the lasso curls around it. The arrangement is stabilized through the actin monomers that have contacts with both FH2 domains. From a different crystal structure, three actin molecules are arranged through a two-fold screw axis and in contact with a dimer of FH2.

Formin also contains a dimerization domain (DD) and at the very C-terminus a small region called FSI (Formin Spire Interaction) that interacts with the KIND domain of Spire.

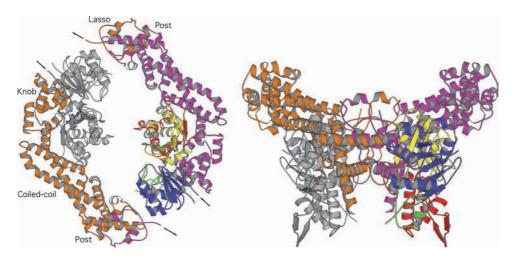


Fig. 15.6 ■ The dimeric structure of the FH2 domain of formin in two perpendicular views. The structure can be divided into five domains or regions: lasso, linker, knob, coiled-coil and post (PDB: 4EAH). The lasso is a rope-like structure that coils around the post part of the other subunit. The linker is disordered and its position indicated by arrows. One of the two actin molecules is shown in the same domain colors as Figure 15.2.

15.1.3 Actin Fiber Nucleation

Most actin monomers in solution are in complex with profilin. Profilin has the role of a nucleotide exchange factor, but also assists in the nucleation and elongation of actin fibers. The nucleation of an actin fiber needs three actin monomers to bind stably to each other. Several protein complexes can perform nucleation of actin fibers, but a central one is the Spire-formin couple.

The Spire protein can both bind to the membrane and dimerize through the FYVE domain. The dimeric Spire can then loosely bind up to eight actin monomers to its WH2 repeats, which are subsequently able to nucleate and form a normal actin fiber. For efficient nucleation, interplay with formin is required. A C-terminal region of formin called FSI (Formin Spire Interaction) can bind to the N-terminal KIND domain of Spire (Figure 15.7). However, most actin monomers are bound to profilin, inhibiting fiber growth, but profilin has a surface with affinity for polyproline structures and can therefore bind to the FH1 domain of formin. The FH2 domain of formin dimers in complex with two or three actin molecules can form a nucleus for fiber growth.

15.1.4 Proteins Capping or Severing Actin Fibers

15.1.4.1 Gelsolin

Gelsolin belongs to a family of proteins severing and capping actin filaments in eukaryotes. One of its physiological functions is to prevent formation of actin fibers in plasma, but it

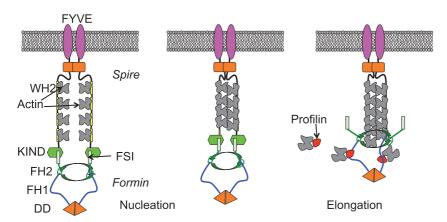


Fig. 15.7 ■ A model for the nucleation and elongation of actin fibers by the proteins Spire and formin. The Spire protein binds in dimeric form to the membrane. Its four WH2 regions can each bind an actin monomer (the pointed end is up and the barbed end is down). The monomers can form a nucleus of an actin fiber. Formin can form an actin nucleus but also participate in the elongation. Complexes of actin and profilin bind to the polyprolin sequences of the FH1 domain and are brought into the growing fiber by the circular FH2 dimer.

is also found inside cells regulating the actin fibers of the cytoskeleton. It is uncommon for a protein to be used both inside and outside the cell. Gelsolin can also nucleate fiber formation.

Gelsolin has six domains with similar folds. Each domain has a mixed sheet with helices on one side. In the complete molecule, domains 1-3 are pairwise similar to domains 4-6, suggesting that the molecule evolved first through a triplication of an ancestral gene, followed by a duplication of this three-domain gene. Domains 1 and 3 form a continuous sheet, as do domains 4 and 6. The different domains have different functions in the regulation of actin polymerization. Domain 1 is mainly responsible for severing fibers.

Crystal structures are known for a single gelsolin domain in complex with actin, a complex with domains 1-3 and a complex with domains 4-6. In the complexes, the gelsolin domains 1 or 4 binds to the target-binding cleft in the same way as profilin. A helix is bound in the same position as the helices in WH2 domains (Figure 15.4). The structure of domains 1-3 or 4-6 with actin shows a large rearrangement of the domains. The continuous sheets formed between domains 1 and 3 and between 4 and 6 in gelsolin are broken, to release the actin-binding surface. In both domain 1 and 4, an aspartate side chain is interacting with a lysine side chain in domains 3 and 6, respectively, to stabilize the interaction in the inactivated protein. Upon calcium binding, this aspartate side chain binds to the ion, the salt bridge is broken and the contact between the domains is weakened, which allows the actin-binding surface to bind to actin (Figures 15.8 and 15.9).

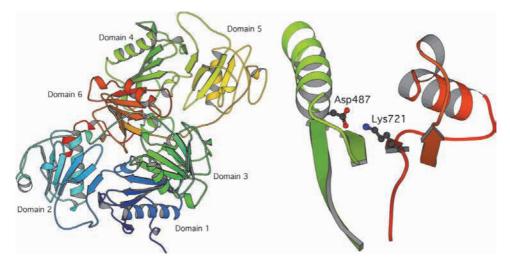


Fig. 15.8 ■ Left: The structure of plasma gelsolin. In this conformation, all actin-binding surfaces are shielded. Controlled by Ca²⁺ ions, the contacts in the complex can be disrupted and the actinbinding surfaces become exposed. Right: A detail of the interaction between domains 4 and 6 where a salt bridge between Asp487 (domain 4) and Lys721 (domain 6) is stabilizing the contact (PDB: 1D0N).

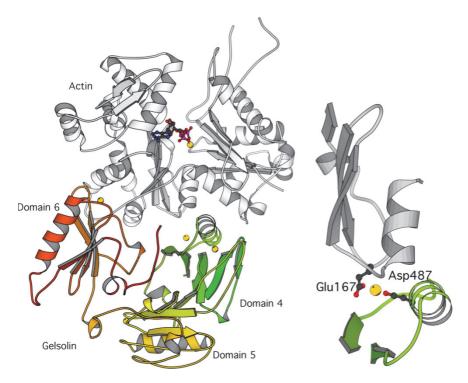


Fig. 15.9 ■ *Left*: The structure of the complex between actin (grey) and the gelsolin domains 4–6. Note that the continuous sheet between domains 4 and 6 is broken in the complex, probably due to activation by calcium ions (yellow). Right: A detail of the contact surface where Asp487 in domain 4 of gelsolin binds to a calcium ion as well as Glu167 from actin (PDB: 1H1V).

15.1.5 Actin Attachment and Cross-linking

$15.1.5.1 \ Arp2/3 \ complex$

The leading edge of the motile eukaryotic cell is pushed by a growing actin network (Figure 15.10). New actin filaments can grow as branches from other filaments. Many proteins are used to control the sites of growth and crosslinking of these actin fibers. A protein complex called Arp2/3 (actin related proteins 2 and 3) generates the branches in these actin networks. At the same time, proteins like cofilin stimulate the depolymerization of the actin network by binding at the barbed end.

Arp2/3 is a complex of seven protein subunits (Figure 15.11). Arp2 and Arp3 are similar to actin and need ATP for their activity (Figure 15.12). Compared to actin, some loops are slightly longer. Arp2 and Arp3 form the starting point at the pointed end for the polymerization of a new actin fiber. The rest of the complex is involved in binding to the actin fiber and controls the nucleation. The complex is inactive until it is bound to an actin fiber and stimulated by nucleation promoting factors (NPF) such as WASP proteins and cortactin. These two proteins work synergistically. Cortactin binds to actin filaments

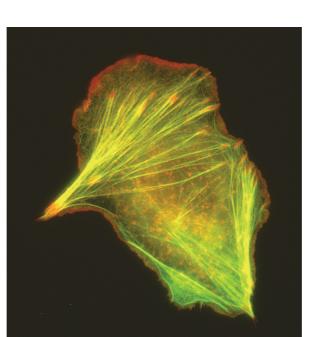


Fig. 15.10 ■ A motile cell. The picture shows a cultured tumor cell (melanoma). The colored fibers (stress fibers) represent a system of protein polymers belonging to the microfilament system built of actin and involved in generating force for cell motility and migration. Hundreds of protein molecules comprise this system. This cell had been transfected with a genetic construct leading to the expression of an actin regulatory protein tropomyosin (TM5, tagged with green fluorescent protein, GFP). Actin was visualized using rhodamine phalloidin (red). Depending on the balance between GFP-TM5 green and actin red, the actin containing filamentous structures differ in color from red through yellow to green. The red blobs at the ends of stress fibers are actin at adhesion sites from which tropomyosin is excluded. Photograph by Louise Bertilsson, kindly provided by Uno Lindberg.

through a 6.5 times repeated sequence of 37 amino acid residues. WASP has two regions, a central amphipathic helix and a C-terminal acidic motif, that interact with Arp2/3. The N-terminal region has a WH2 domain (for this protein called V) that binds actin monomers. In the activated complex, Arp2 moves about 25 Å to form a dimer with Arp3 like a short pitch of an actin filament providing a template for nucleation of the new filament. Here, the WH2 element of WASP can contribute actin monomers.

Another common fold is the propeller domain of ARPC1, which is very similar to the β subunits of the trimeric G-proteins involved in cell signaling (Section 14.4). Of the remaining subunits ARPC2 and ARPC4 have the same topology with an antiparallel sheet of the meander type and a long C-terminal helix. p34 has two copies of this fold. The ARPC3 and ARPC5 proteins are both helical.

A protein called GMF (glial maturation factor) inhibits the nucleation and can disassemble branches of actin filaments by binding specifically to the barbed end of



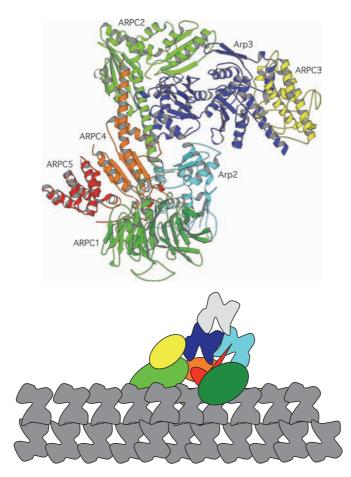


Fig. 15.11 ■ Top: The Arp2/3 complex (PDB: 1K8K). Bottom: A schematic model of how the complex (in the colors of the complex above) introduces a branch of actin fibers (gray). An actin monomer is added (light gray) to the growing branch.

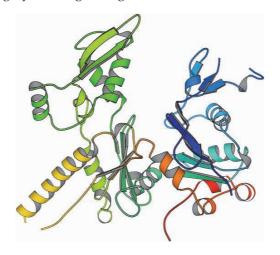


Fig. 15.12 ■ The actin-like Arp3 subunit of the Arp2/3 complex in the "standard" orientation. The nucleotide-binding cleft is more open than in actin.

Arp2, in a manner similar to the way cofilin binds to actin filaments. GMF also interacts with other proteins in the Arp2/3 complex.

15.2 Myosin and Muscle Function

15.2.1 Muscle Architecture

Muscle cells contain bundles of contractile protein fibers called myofibrils. Along the myofibrils one can see (with an electron microscope) a repetitive pattern of light and dark bands and discs (Figure 15.13). The dark bands consist of thick filaments and the light bands of thin filaments. The discs are the structures to which the thin filaments are attached. One repeated unit along the fiber is called a sarcomere, and it begins and ends at the Z-disc. About 300 myosin molecules, that are cross-linked at the M-band, form the thick filaments. The main component of the thin filament is an actin fiber. The thin filaments are partly embedded between the thick filaments in a regular arrangement (Figure 15.14).

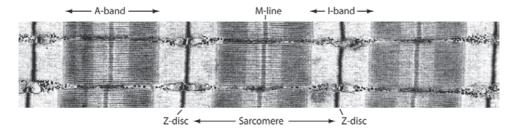


Fig. 15.13 ■ Electron micrograph of a muscle fiber and a schematic illustration of the arrangement of actin and myosin in the sarcomere. The dark A-band consists of thick filaments and the light I-band of thin filaments attached to the Z-disc. The thick filaments are cross-linked in the M-line. The thick and thin filaments overlap in the darkest region. (Courtesy of Dr Roger Craig, University of Massachusetts.)

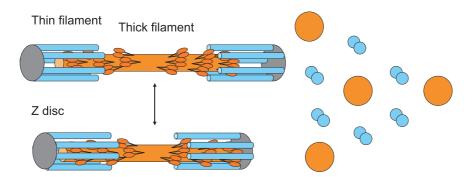


Fig. 15.14 ■ Schematic drawing of a piece of a sarcomere in the extended and contracted form. To the right is a schematic view of the arrangement of thick (orange) and thin (blue) filaments in the region where they interact.

The Z-disc is composed of a multitude of proteins, where α -actinin crosslinks the barbed ends of actin fibers from two sarcomeres, meeting at the Z-disc. Another of these proteins is titin, an exceptionally long protein with upto 35 000 amino acid residues and a molecular mass of up to 4 MDa. It extends from the M-band to the Z-disc and may have a role in maintaining the length of the sarcomere and the position of the M-band in the middle of the sarcomere.

Muscle contraction is generated by the motor protein myosin that makes the filaments slide against each other to shorten the sarcomere. The thin and thick filaments are cross-linked by protein-protein interactions. Sliding is caused by moving the thick filaments relative to the thin filaments towards the Z-disc in forming and breaking the cross-links between the thick and thin filaments (Figure 15.14).

15.2.2 Proteins Binding to Actin Fibers in Thin Filaments. Tropomyosin and Troponin

The thin filaments are associated with two proteins that are involved in the calcium-dependent regulation of contraction. One of them is tropomyosin, which is a very elongated molecule. In mammals there are four genes for tropomyosin, but these can generate 40 different isoforms. There are two main forms, muscle and non-muscle tropomyosin. Tropomyosin is built up of two identical α -helical chains that form a 385 Å long coiled-coil head-to-tail dimer (Section 3.3.2). Its sequence has about 40 segments of a heptad repeat, abcdefg, where a and d are mostly non-polar as in other coiled-coils. Tropomyosin can also be divided into seven pseudorepeating units, which form a continuous structure wound around the actin helix in the thin filament (Figure 15.15). Each dimer binds to seven pairs of actin monomers. Naturally, there are two tropomyosin dimers bound to the actin filament, one to each side.

Tropomyosin, which is generally negatively charged, interacts with a positively charged groove of actin. There are two positions for tropomyosin on the actin filament. One of them partly blocks the myosin binding sites on the thin filament.

Troponin is important for the calcium ion sensitivity of muscle. Troponin is a complex of three chains, C, I and T. Troponin C is a calcium-binding protein with four EF-hands. It binds to the other proteins and forms a regulatory head. The other two molecules form long helices (Figure 15.16). One troponin complex is associated with every tropomyosin molecule and controls the activity by moving the tropomyosin molecule, thereby removing the steric blocking of the binding site for myosin on the actin fiber. The stoichiometry of tropomyosin:troponin:actin is 1:1:7. Tropomyosin functions as a molecular ruler placing the Ca²⁺ controlled troponin at regular intervals. At low calcium concentration, troponin holds tropomyosin in a closed state, but at increased calcium the tropomyosin position is altered to make myosin able to bind.

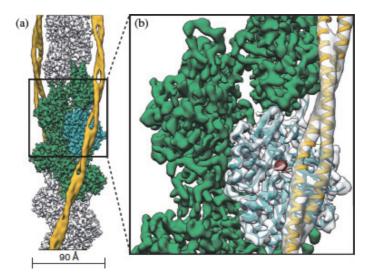


Fig. 15.15 ■ The structure of actin at 3.7 Å and tropomyosin at 6.5 Å resolution obtained by cryo-EM. (Reproduced with permission from von der Ecken J et al. (2015) Structure of the F-actin — tropomyosin complex. Nature 519: 114–117. Copyright Macmillan Publishers Ltd.)

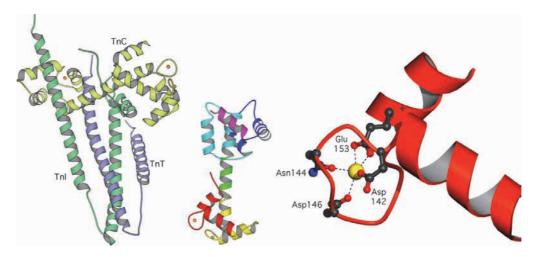


Fig. 15.16 ■ Left: The troponin complex. Troponin I and the C-terminal part of troponin T look like chopsticks formed by two helices (PDB: 1YTZ). TnC is located as a hand holding the chopsticks. Middle: Troponin C is a typical EF-hand protein with two pairs of calcium ion binding EF-hand motifs connected by a long helix. The EF-hands are colored blue, cyan, yellow and red. In this crystal structure, only the C-terminal pair of EF-hands binds calcium (PDB: 5TNC). Right: A close-up view of the binding of the calcium ion. Oxygen atoms from Asp, Asn and Glu side chains coordinate the ion together with a carbonyl oxygen and a water molecule (not shown).

15.2.3 Myosin

Myosins are ATPases that are able to move along fibers; they are mechanochemical enzymes and motor proteins. They use the chemical energy in ATP to perform active work on a molecular level, resulting in large-scale motion. There are several types of myosins. The myosin in the thick filaments of muscle is called myosin II. Myosins I and V are associated with membranes and are part of the active cytoskeleton of cells (Table 15.2).

Each myosin molecule has a head (the motor domain), a neck and a tail (Figure 15.17). The main component of myosin is the heavy chain. Its tail is formed by the C-terminal part, which is responsible for dimerization of the molecule. In the case of myosins I and V, it controls the association with membranes. A large portion of the tail has a repeat of 28 amino acids. This is in agreement with the proposed conformation of the tail: the two heavy chains form a coiled-coil of two very long a helices winding around each other. Myosins V and VI have C-terminal domains that are responsible for binding of cargo for cellular transport (vesicles or organelles).

All myosin heavy chains are associated with calcium-binding proteins of the EF-hand type. These are found in the neck region. In myosin II, the heavy chains are associated

| Type | Molecular Weight of Heavy Chain (kDa) | Main Function |
|------|--|--|
| I | 110–150 | Membrane binding. |
| II | 220 | Filament sliding in muscle. |
| V | 170–220 | Vesicle transport. |
| VI | 140 | Transport of endocytotic vesicles, moves towards pointed end of the actin fiber. |

TABLE 15.2 Some Types of Myosins and Their Functions

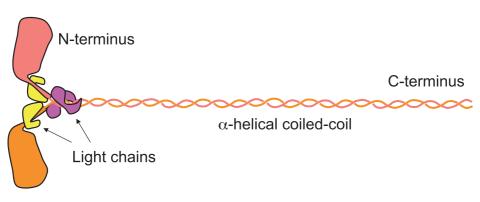


Fig. 15.17 ■ The head, neck and tail of a myosin II molecule from left to right. There are two heavy chains and two pairs of light chains.

with the regulatory light-chain (RLC) and the essential light chain (ELC) proteins, while myosins I, V and VI are associated with calmodulin.

15.2.3.1 Structure of crossbridge

The N-terminal fragment of myosin II called S1 (Figure 15.18), contains the head and neck regions. The C-terminal tail region has about 1000 amino acid residues. The neck consists of the C-terminal part of S1 with the two light chains. The neck has a long and bent helix, with the light chains arranged around this helix. The light chains are related to calmodulin and troponin C with two globular domains, each formed by two EF-hands and connected by a flexible and partly helical segment. Only some of the EF-hand motifs in RLC and ELC are able to bind calcium ions.

The head of myosin is a complex structure with a number of domains. At the N-terminus, there is a small antiparallel β -sheet or β -barrel (the N-terminal domain, NTD). The central part of the head is a β -sheet that is formed by strands that originate from three different segments of the chain. This sheet is mainly parallel. The connections between the strands are mainly helical. Some of these helices form two separate domains, called the upper and lower 50 K domains. The C-terminal part of the head forms a small domain called the converter domain. The neck or lever arm and the tail are directly connected to the converter domain.

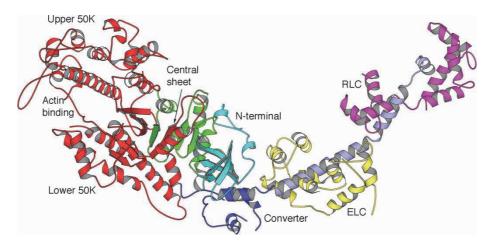


Fig. 15.18 Chicken myosin S1 fragment (myosin head and neck region) showing the N-terminal part of the heavy chain and the two light chains (PDB: 2MYS). The head has five domains: the N-terminal barrel (cyan), the upper and lower 50 K domains (red) and the converter domain (blue) that leads to the C-terminal helix (light blue). The central part of the head is formed by a sheet (green). The ATP molecule binds to this sheet. The neck or lever arm consists of ELC (essential light chain, yellow) and RLC (regulatory light chain, purple) bound to the C-terminal helix. Both ELC and RLC consist of two lobes (EF hands) connected with a linker (Figure 15.16). ELC and RLC define the length of the lever arm; other myosins have up to six calmodulin proteins bound, resulting in larger movements of the head relative to tail.

15.2.3.2 Relation between myosin structure and the G-proteins

The central strand has a sequence pattern found in many ATPases: GXXXXGKS/T, the P-loop (Section 8.3). The ATP molecule binds at this position, similar to GTP binding by the P-loop in G-proteins (Figure 15.19).

The topology of the central sheet in myosins is not identical to that in the G-proteins, but it is nevertheless possible that myosins kinesins and the G-proteins are distantly related by evolution. In addition to the similarity of their P-loops, the four central occur in the same order (strands 2, 3, 1, 4, in Ras; 4, 6, 3, 7 in myosin; 6, 7, 3, 8 in kinesis). The other strands of the sheets come from different segments.

15.2.3.3 Mechanism of myosin II in muscle contraction

In muscle contraction, the binding, hydrolysis and release of ATP controls the binding of the cross-bridge to the actin filament, the power stroke and the binding to a new position of the filament. It has not been possible to crystallize complexes of actin fibers with myosin, and our understanding of the mechanism is based on crystallographic studies of myosin fragments (heads and complete cross-bridges) fitted into cryo-electron microscopy maps



Fig. 15.19 A comparison of the sheet and the P-loop in the G-protein Ras (*left*), myosin (from slime mold, PDB: 1MMA, *middle*), and kinesin (PDB: 1BG2, *right*). The order of the β-strands in the sequence is indicated. Some of the strands are in the same order in the sequence (in red). The upper and lower 50 K domains in myosin are long insertions between strands 5 and 6, and between 6 and 7.

of actin fibers in complex with myosin heads and also with tropomyosin (Figure 15.20). The resolution of these micrographs allows a good definition of the interacting molecules.

There are two conformations that illustrate the position of the lever arm, exemplified by the original structures of chicken skeletal muscle myosin and chicken smooth muscle myosin (Figure 15.21). These structures may represent what has been called a post-rigor

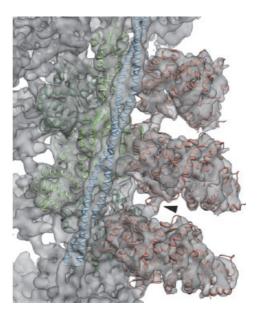


Fig. 15.20 ■ Fitting of actin (green), tropomyosin (blue) and myosin heads (red) into a cryo-EM density at 8 Å resolution. (Reproduced with permission from Behrmann et al. (2012) Structure of the rigor actin-tropomyosin-myosin complex. Cell 150: 327–338. Copyright (2012) Elsevier.)

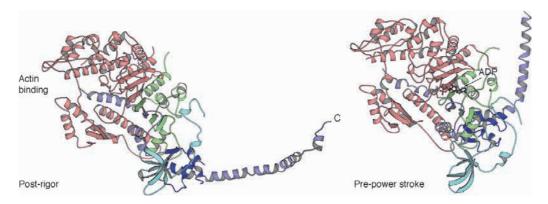


Fig. 15.21 ■ The myosin S1 fragment at two extreme positions of the neck. *Right*: Chicken smooth (ADP and AlF₄, PDB: 1BR1), prepower stroke state. Left: Chicken skeletal (no nucleotide bound, PDB: 2MYS), post-rigor state. The head is in the same orientation and the large variation in the orientation of the neck is obvious. The post-rigor state roughly corresponds to the orientation of the arm at the end of the cycle, while the prepower stroke state is close to the state when the head is binding to a new position of the actin fiber.

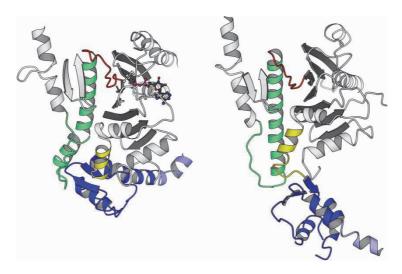


Fig. 15.22 A more detailed view showing conformational changes in the myosin head leading to changes in the relative orientation of head and tail. The central sheet is in dark grey. Right: Chicken skeletal and *left*: Chicken smooth muscle. The subdomains are connected by segments that vary in conformation: switch II (red), the SH1 helix (yellow) and the relay helix (green). The conformation of switch II is different, coupled to changes in the relay helix and the converter domain (blue). These segments are sensitive to the presence of ADP or ATP, but also to binding to the actin fiber.

state and a prepower stroke, respectively. The post-rigor state and the prepower stroke state are close to the extreme conformations at the beginning and the end of the conformational changes of the molecule. In the post-rigor state, the arm is found at the position after the movement (the power stroke) has occurred, but the head is no longer bound to the actin fiber. In the prepower stroke state, the arm is "reprimed" and is at the position when the myosin head is rebinding to the actin fiber after ATP cleavage.

The large movement of the lever arm is controlled by a number of flexible parts of the head. The switch 1 and 2 loops have important functional roles, as in G-proteins (Section 8.3). The movement of the lever arm is controlled by the converter domain, which is rigidly connected to the neck. The orientation of the converter domain is in turn controlled by the relay helix, which is connected to the switch 2 loop (Figure 15.22). In the absence of a bound nucleotide (corresponding to the rigor state), switch 2 is at some distance from the nucleotide-binding site and the relay helix is straight. ATP binding allows switch 2 to move closer to the nucleotide-binding site and induce a kink in the relay helix when the converter rotates to its position before the power stroke ("repriming").

The affinity of the myosin head for the actin fiber is controlled by relative movements of the upper and lower 50 K domains (Figure 15.23). Strong binding to actin corresponds to a closing of the gap between these domains at the actin fiber. ATP binding probably leads to changes in switch 1, that in turn leads to opening of the gap and loss of binding. The movements of switches 1 and 2 lead to an active enzyme and ATP hydrolysis. Closing of the gap opens up the entrance to the nucleotide-binding site, allowing the release of ADP after hydrolysis.

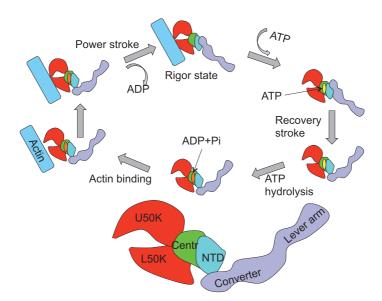


Fig. 15.23 ■ The steps in muscle contraction. *Top*: Rigor state, followed by ATP binding and hydrolysis, actin binding and power stroke. (After Behrmann *et al.* 2012. Structure of the rigor actin-tropomyosin-complex. *Cell* **150**: 327–338.) *Bottom*: The head domains of myosin.

15.2.4 Myosin II in Muscle Contraction

Myosin II is found in animals as well as simple organisms like yeast. It is a major part of thick filaments of muscle fibers, but it is also found in other types of cells. Myosin II is composed of six chains: two copies of the heavy chain of about 1950 amino acids and two copies each of RLC and ELC.

When muscles contract, the myosin molecule moves the filaments relative to each other. The energy required for this work comes from cleavage of ATP in myosin. The thick filament is bipolar and the myosin heads form a helical array in both ends of the filament. The heads and necks of the myosin molecules form cross-bridges by binding to the actin fiber of the thin filaments. ATP hydrolysis leads to a relatively large conformational change in the neck of the myosin molecule. The neck acts as a lever arm ("swinging lever arm model") and causes the filaments to slide relative to each other, moving the thick filament towards the "+" ends of the actin fibers or the Z-discs and shortening the sarcomere. The myosin head is then released from the actin fiber and is able to bind at another position to repeat the cycle. With the large number of heads along a thick filament, some binding and some detached, there is little risk of slippage.

In the absence of ATP, myosin binds strongly to the actin fiber ("rigor" state). According to one hypothesis, the myosin cycle involves the following steps:

(1) ATP binding leads to a conformational change breaking the interaction between myosin and actin (post-rigor state).

- (2) ATP hydrolysis leads to a conformational change allowing myosin to bind at a new position on the actin fiber. ATP hydrolysis probably occurs before the myosin head is attached to actin.
- (3) The attachment to actin leads to a closing of the gap between the upper and lower 50 K domains. This leads to the opening of the ATP/ADP-binding pocket.
- (4) The phosphate leaves myosin causing another conformational change.
- (5) When ADP is released, myosin returns to its resting or "rigor" state, ready to restart from step 1. This leads to the movement of the fibers ("power stroke").

15.2.4.1 Actin and myosin in transport

Actin filaments are also involved in transport and in developing cell polarity. While microtubules are involved in long distance transport, actin and myosin (type VI) transport cellular substances like mitochondria and mRNA to suitable nearby locations. Viral components can also hijack the actin transport system for rapid assembly of new virions.

15.3 Microtubule

15.3.1 Tubulin Structure and Function

Microtubules are tubular structures that have two functions in cells. They are used for intracellular transport of vesicles in the cytoplasm and they are also involved in cell division, where they separate the replicated chromosomes. Microtubules are mainly composed of tubulin, a protein with two subunits, α and β , forming heterodimers. These dimers are arranged head-to-tail in protofilaments. Thirteen protofilaments interact side by side in a controlled way to form the microtubule with a diameter of around 25 nm (Figure 15.24). This cylinder has a stiffer structure than the actin microfilaments and is used for functions that need mechanical stability. Microtubules grow and disassemble in a way similar to actin fibers, processes controlled by other proteins.

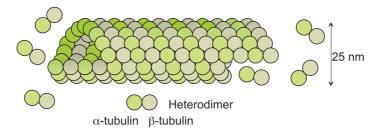


Fig. 15.24 \blacksquare Microtubules are stiff molecular tracks for transport built of tubulin α and β heterodimers.

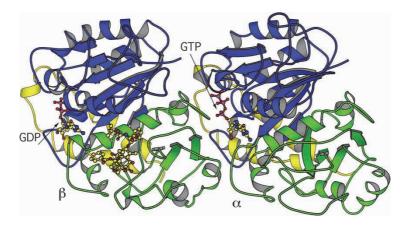


Fig. 15.25 ■ Structure of a tubulin αβ dimer. The Rossmann fold domain (blue) is at the top. The second domain is in green and the C-terminal helices in yellow. A drug molecule bound to the β protein was included in the structure determination (PDB: 1TUB).

The α - and β -tubulin monomers both bind GTP, but only the β subunits have GTPase activity. In the growing end of a microtubule, all monomers will contain GTP, but further along in the structure GTP has been degraded to GDP in the β subunits. The GTP content in the ends of the microtubule defines the state of it as elongating or shrinking.

The α and β subunits have very similar compact conformation (Figure 15.25) formed by three domains. The first, GTP-binding domain, has the classical Rossmann fold of six parallel β-strands connected by helices similar to dehydrogenases. The second domain is a mixed sheet with helices on both sides. The third domain has two helices packed against the first domain. The topology of the GTP-binding domain is similar, but not identical to that of the G-proteins, and there is no GXXXXGKT/S sequence in the loop that corresponds to the P-loop and that interacts with the nucleotide phosphates.

The difference in GTPase activity of the α and β subunits is understood from the structure. In the dimer, the GTP site of the α subunit is blocked by the β subunit. However, the site in the β subunit is accessible to water, allowing hydrolysis to be stimulated by the α subunit.

Initiation of the growth of microtubules depends on a third kind of tubulin, γ-tubulin, with a similar structure and GTPase activity as other tubulins. Some proteins stabilize the filamentous form of tubulin, like microtubule-associated proteins (MAPs), while others have a destabilizing function, like proteins of the stathmin family or the drug colchicine.

15.3.2 Microtubule-associated Motor Proteins

The transport of vesicles along microtubules uses two types of force-producing proteins dyneins and kinesins. Kinesins are motor proteins with properties similar to myosins and with a common evolutionary background. A more complex relative is the motor protein dynein. It is primarily composed of a ring of six AAA+ domains in the same polypeptide and a number of additional subunits. The dynein complex also transports cargo along the microtubule.

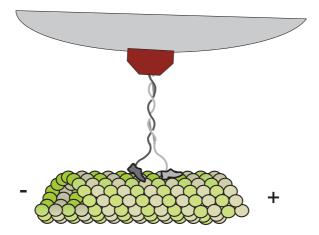


Fig. 15.26 ■ The kinesin molecule transports large pieces of cargo by performing something like a walk on the tubulin track with its motor domains, here symbolized by feet.

The kinesin family is subdivided into three classes: N-, C- and M-types, depending on the position of their motor domain (N-terminal, C-terminal, middle). N-type kinesins are used for transport along microtubules. They use ATP hydrolysis to generate the force used to move organelles or chromosomes. Most kinesins are dimers of two identical heavy chains of about 1000 amino acids. The N-type (like myosin) has an N-terminal motor domain (the head) and a long helical region responsible for dimerization (the stalk). The cargo to be transported is attached through light chains that are associated with the C-terminal end of the tail helix (Figure 15.26).

15.3.2.1 Kinesin structure and function

The motor domain of N-type kinesins is similar to the myosin head, although much smaller (compare Figure 15.27 with Figure 15.19). They have a mainly parallel sheet with helices on both sides. The difference in size is chiefly due to large insertions at a few positions in the myosin structure that forms the actin-binding cleft. A comparison of the sheets of myosin, kinesin and a G-protein is shown in Figure 15.19. The order of the strands surrounding the phosphate-binding loop is the same.

The two kinesin heads bind to the microtubule one at a time. At the movement, one head swings over the other to attach to a new site (Figure 15.28). In this way, the kinesin dimer can pull an object, for example, an organelle, attached to the stalk part of the kinesin. The conformational changes in the head domain may be similar to those in the myosin head. The neck linker connecting the coiled-coil to the motor domain is a conserved segment of 15 amino acids, and is essential in this function. This segment is docked to the core of the head when ATP analogs are bound to the head, but disordered in other structures. Part of it is fixed to the head through a β-strand at the N-terminus.

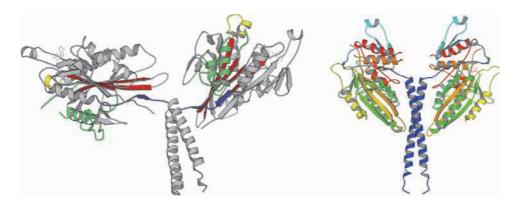


Fig. 15.27 ■ *Left*: A kinesin dimer with a helical neck region (PDB: 3KIN). The coiled-coil stalk of about 600 amino acids begins at the end of the helical neck. The central part of the sheet (red, see Figure 15.19), the switch I region (yellow), switch II and the "switch II cluster" binding to microtubules (green), and the linker between the motor head and the neck (blue). Note that the dimeric motor domains do not obey two-fold symmetry. *Right*: An ncd dimer (a C-type kinesin, PDB: 2NCD). There is no linker between the motor domain and the stalk.

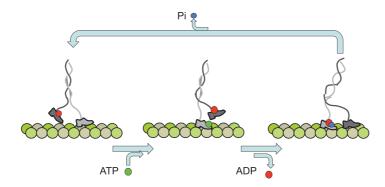


Fig. 15.28 ■ The steps involved in the walk of the kinesin "shoes" on the microtubulin track.

A related motor is the protein ncd (non-claret disjunctional), which is active in chromosome separation. This protein belongs to the C-type kinesins, where the motor domain is found at the C-terminal end of the protein. It is relatively similar to N-kinesin, but it acts in the opposite direction, moving towards the "–" end of the microtubule. This difference is probably connected to the position of the motor domain in the chain. There is no neck linker in this type of kinesin.

The third kind of kinesins, the M-type or KinI kinesins, have a similar motor domain, but do not transport substances along microtubules. Their function is to bind and depolymerize microtubules. The structure of the catalytic core shows that it is similar to that in other kinesins. As in the case with other kinesins, it is difficult to understand the mechanisms of its action as long as there are no structures of kinesin-tubulin complexes.

For Further Reading

Original Articles

- Behrmann E, Müller M, Penczek PA, et al. (2012) Structure of the rigor actin-tropomyosin-myosin complex. Cell 150: 327-338.
- von der Ecken J, Müller M, Lehman W, et al. (2015) Structure of the F-actin Tropomyosin complex. *Nature* **519**: 114–117.
- Kabsch W, Mannherz, HG, Suck D, et al. (1990) Atomic structure of the actin: DNase 1 complex. Nature 347, 37-44.
- Nogales E, Wolf SG, Downing KH. (1998) Structure of the αβ tubulin dimer by electron crystallography. Nature 391, 199-203.
- Rayment I, Rypniewski WR, Schmidt-Bäse K, et al. (1993) Three-dimensional structure of myosin subfragment-1: A molecular motor. Science 261, 50–58.
- Robinson RC, Turbedsky K, Kaiser DA, et al. (2001) Crystal structure of arp2/3 complex. Science **294**(5547), 1679–1684.
- Rouiller I, Xu X-P, Amann KJ, et al. (2008) The structural basis of filament branching by the Arp2/3 complex. J Cell Biol 180: 887-895.
- Thompson ME, Heimsath EG, Gauvin TJ, et al. (2013) FMNL3 FH2-actin structure gives insight into formin-mediated actin nucleation and elongation. Nat Struct Mol Biol 20: 111-118.

Reviews

- Dietrich S, Weiss S, Pleiser S, Kerkhoff E. (2013) Structural and functional insights into the Spir/ formin actin nucleator complex. Biol Chem 394: 1649–1660.
- Dominguez R, Holmes KC. (2011) Actin structure and function. Ann Rev Biophys 40: 169–186.
- Firat-Karalar EN, Welch MD. (2011) New mechanisms and functions of actin nucleation. Curr Opin Cell Biol 23: 4-13.
- Moore JR, Campbell SG, Lehman W. (2016) Structural determinants of muscle thin filament cooperativity. Arch Biochem Biophys 594: 8–17.
- Schmidt H. (2015) Dynein motors: How AAA+ ring opening and closing coordinates microtubule binding and linker movement. *Bioessays* **37**: 532–543.
- Wang W, Cao L, Wang C, et al. (2015) Kinesin, 30 years later: Recent insights from structural studies. Prot Sci 24: 1047–1056.

Structural Aspects of Cell-Cell Interactions

16.1 Proteins in the Extracellular Matrix

In animal tissues, cells are organized within a matrix called the extracellular matrix (ECM). The main components of the extracellular matrix are collagen fibers, proteoglycans (Chapter 7) and various matrix proteins with the capacity to bind to other components. Depending on the tissue, the composition of the matrix differs to confer each tissue the right properties. The components of the extracellular matrix are synthesized by the cells in the tissue and are continuously renewed.

16.1.1 Collagen Fibers

Collagens are fibrous proteins that are the main components of many tissues, e.g. tendons, ligaments, skin, blood vessels and dentin in teeth. It is the most abundant protein in our bodies. Its function is structural, but many other proteins bind to collagens in tissues and the extracellular matrix.

There are 44 genes for collagens in the human genome. They are coding for proteins with between 700 and 3000 amino acid residues. These combine to form many different types of collagen fibrils with properties suitable for different types of tissue. Collagens are synthesized as precursor molecules with a fiber-forming central part of about 1000 amino acid residues flanked by N- and C-terminal domains. These domains are important for the assembly of the trimeric tropocollagen and are removed by proteolysis. In tropocollagen, the chains wind around each other into a triple helix, to form fibrils where packages of collagen triple helices are staggered to form long fibers. The triple helices are covalently linked through lysine side chains. The lysine residues are first oxidized to

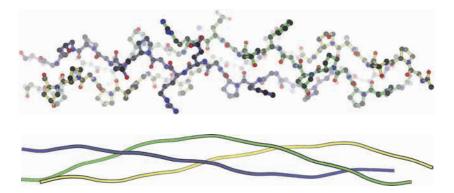


Fig. 16.1 ■ Structure of a segment of tropocollagen, a triple helix. Top: Ball-and-stick model. Bottom: Rope model of the three chains using the same colors as in the top drawing. The three chains primarily have a repeated sequence of proline-hydroxyproline-glycine. The glycine residues allow the chains to come into close contact and form interchain hydrogen bonds (PDB: 1Q7D).

aldehydes by the extracellular enzyme lysine oxidase, and subsequently the aldehyde group reacts with lysines in neighboring triple helices forming crosslinks.

The sequence of the collagen chains is highly repetitive with the sequence X-Y-Gly, where residues X and Y often are proline or hydroxyproline residues. Hydroxyprolines are the result of a post-translational modification of proline residues catalyzed by a special enzyme. The collagen chain forms a superhelix of three stands (see Section 2.3.1.5). The chains in the triple helix are connected by hydrogen bonds, and these can form because the prevalence of glycine residues allows a close packing of the chains (Figure 16.1).

16.1.2 Fibronectin

The extracellular matrix contains aggregates of the glycosylated protein fibronectin, and cells attach to the extracellular matrix mainly through binding to this protein. Fibronectin is also found in plasma, where it exists as dimers of two chains that are linked by C-terminal disulfides. In the extracellular matrix, each chain binds to several other molecules, like collagen, heparan sulfate, fibrin or integrins.

The single fibronectin gene has 47 exons, which produces several different proteins due to different splicing. Fibronectin has a modular structure of 220 kDa with six domains, where each consists of a number of modules (Figure 16.2). The modules are of three types, I, II and III, and several copies of each module are found in a pattern that varies slightly between different fibronectin molecules. Similar modules have been identified in many other proteins based on sequence similarity.

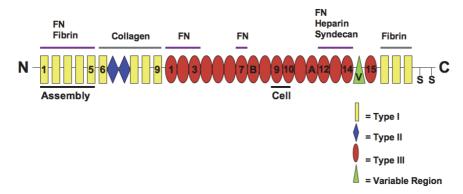


Fig. 16.2 ■ Schematic structure of a monomer of human fibronectin. It can be divided into six regions or domains according to its binding properties: regions 1, 3, 4, 5 interact with other fibronectins (FN); 2 with collagen; 1 and 5 with heparin; and 1 and 6 with fibrin. Each region is composed of a number of modules with different color as indicated. The modules of the three kinds (I, II, III) are numbered separately and sequentially from the N-terminus as indicated. One chain is linked to the next one by disulfide bonds close to the C-terminus. The illustration was obtained from Wikipedia. Author: AllWorthLettingGo.

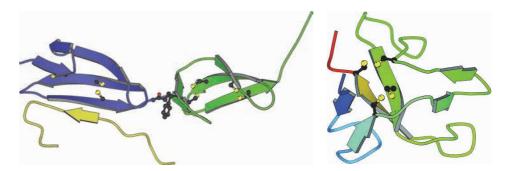


Fig. 16.3 ■ Fibronectin type I and II modules. Left: Type I modules 8 (blue) and 9 (green) from human fibronectin with a bound peptide from collagen (yellow). There are disulfide bonds between strands 1 and 4 and between strands 4 and 5 in each module. Two side chains from module 9 that bind in a pocket in module 8 are shown. Note that the collagen segment is extended when bound to the fibronectin modules (PDB: 3EJH). Right: A type II module (PDB: 2FN2).

The fibronectin type I module has about 45 amino acids. These form a beta hairpin packed on top of a small three-stranded sheet. Such a small structure would in general not be very stable, but two disulfide bonds link the secondary structure elements together to form a stable module.

Fibronectin and integrin receptors are also important in the formation of collagen fibrils in vivo. Collagen interacts with the second domain of fibronectin (Figure 16.3).

The modules in fibronectin, like in other modular proteins, are small but separately folded units. The binding affinity of a single module is weak in most cases. A high affinity binding requires a large area of contact. Therefore, interaction with more than a single module is often required for binding to other molecules, and this means that the relative orientation of the modules may have to be fixed, for example, by direct interaction, as in the case of modules 8 and 9 in fibronectin.

The type II module is also small, with about 45–60 residues. The structure is similar to that of the type I module and is also stabilized by disulfide bridges, but at different positions compared to type I modules.

The fibronectin type III module is a β sandwich similar to immunoglobulin domains. This type of module is present in as much as 2% of all animal proteins, including receptors and extracellular proteins as well as intracellular proteins. Fibronectin type III modules are involved in many types of interactions. In general, the loops connecting the strands of the sandwich determine the specificity of the interactions, primarily to integrins and heparin.

The integrin binding to fibronectin is mainly due to a single module, number 10. It has an RGD sequence motif in one of the loops, which is longer than in other type III modules (Figure 16.4). Integrin binding also depends on domains 8 and 9 and a site, called the synergy site, that has been identified in module 9. The intermodule orientation seems to be important for the interaction with integrins. The structure has been determined for modules 7–10, as well as for many other segments containing several type III modules. The relative position of modules 9 and 10 is unusual in that they have roughly the same orientation.

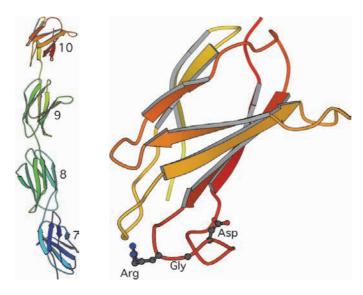


Fig. 16.4 ■ Fibronectin type III modules. *Left*: Modules 7–10 of human fibronectin (PDB: 1FNF). *Right*: Module 10. The side chains of the RGD sequence motif are shown (PDB: 1FNA).

16.2 Linking Cells to the Extracellular Matrix or to Other Cells

Cellular adhesion molecules (CAMs) attach cells to the extracellular matrix or to other cells. The major families of CAMs are shown in Table 16.1. In this chapter, we will discuss only the integrin family.

16.2.1 Integrin Composition and Structure

Integrins are membrane-bound proteins with extracellular regions that bind cells to the extracellular matrix or mediate cell-cell interactions. Integrins are able to perform both inside-out and outside-in signaling. In inside-out signaling, signals from the cytoplasm, caused by activation by other receptors, change the conformation of the integrin from inactive to active to enable it to bind external ligands with high affinity. Ligand binding on the outside of cells can change the conformation of integrin in a way that is sensed in the cytoplasm — outside-in signaling, which can activate the actin polymerization to make the cells migrate.

Integrins are composed of two chains, α and β . The α chain is larger, 150–180 kDa, while the β chain is about 90 kDa. They are cell surface receptors and both the α and β chains have an extracellular part (the ectodomain), a transmembrane helix, and a small intracellular part of about 30 and 50 amino acid residues, respectively. There are at least 18 different α chains and 8 β chains in the human genome. These are combined in different ways and expressed differently depending on the cell type (Table 16.2). The different integrins bind to many ligands, for example, collagens and fibronectin in the extracellular matrix, but also

TABLE 16.1 Families of Cellular Adhesion Molecules. All These Proteins Have a Transmembrane Helix and a Small Intracellular Domain

| Name of Family | Type of Contact | Fold |
|----------------|---|---|
| Cadherins | Cell-cell contacts, same type of cells | Five domains with cadherin fold. Binds to cadherins on other cells |
| Ig-type CAMs | Cell-cell contacts, same type of cells | Several domains with immunoglobulin fold. Binds to CAMs on other cells |
| Integrins | Adhesion to extracellular matrix | Two multidomain chains |
| Selectins | Cell-cell contacts, different types of cells | N-terminal lectin domain (C-type lectin). Binds carbohydrate structures on other cells |

| Beta Chain | Alpha Chain ^a | Examples of Ligands | Sequence Recognized |
|---------------|--------------------------|----------------------------------|---------------------|
| β1 | α1, α2, α10, α11 | Collagens, laminins | GFOGER ^b |
| | α4, α9 | Fibronectin, VCAM-1 | LDV |
| | α5, α8, αV | Fibronectin, vitronectin | RGD |
| | α3, α6, α7 | Laminins | |
| β2 | αL | ICAM-1, ICAM-2 | |
| | αM , αX | Complement C3b, fibrinogen, ICAM | |
| | αD | ICAM-3, VCAM-1 | |
| β3 | αν, αΙΙb | Fibrinogen, many others | RGD |
| $\beta 4^{c}$ | α6 | Laminin | |
| β5 | αV | Vitronectin | RGD |
| β6 | αV | Fibronectin | RGD |
| β7 | α4 | Fibronectin, VCAM-1, MAdCAM-1 | LDV |
| | αΕ | E-cadherin | |
| β8 | αV | Vitronectin | RGD |

TABLE 16.2 Some Integrins and Their Ligands

to proteins on the surface of other cells. The most common feature that is recognized by integrins is the arginine-glycine-aspartate (RGD) sequence, which is found in, for example, fibronectins.

The extracellular part of integrins has a head formed by the N-terminal regions of the α and β subunits. This head is connected to the transmembrane part through a stalk (Figure 16.5). The head portion of the β chain consists of three domains: the I or A domain, the hybrid domain and the PSI (plexin-semaphorin-integrin) domain (Figure 16.6). The hybrid domain is an insertion in the PSI domain; the I domain in turn is an insertion in the hybrid domain. The hybrid domain is a version of the immunoglobulin fold. The I domain is responsible for ligand binding, with a fold similar to the Rossmann fold found in many nucleotide-binding proteins, but the connections between the strands in the N-terminal part of the central sheet are different.

The head portion of the α chain consists of a propeller domain with the same fold as the β chain of the trimeric G-proteins (Figure 16.6). The interaction between the propeller domain and the I domain is also similar to the interactions between the α and β subunits of the trimeric G-proteins. This similarity suggests that the interaction between the propeller

 $[^]a$ The α chains associated with the $\beta2$ integrins as well as the collagen-binding $\alpha1,\,\alpha2,\,\alpha10$ and $\alpha11,$ and αE all have an inserted I domain in their head region.

^b The O in the GFOGER sequence stands for hydroxyproline.

^c The β4 chain has an intracellular part of about 1000 amino acid residues, in contrast to the small intracellular domains of all other integrin chains.

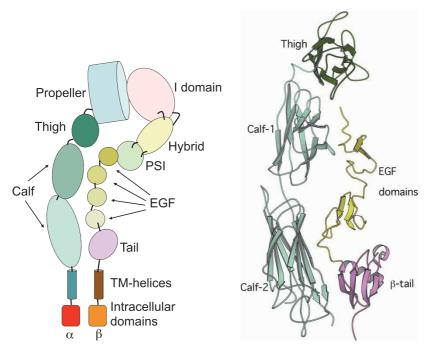


Fig. 16.5 ■ *Left*: A schematic view of an integrin molecule indicating the names of the domains. Right: The stalk of the $\alpha V\beta 3$ integrin. Only two of the EGF domains of the β chain are visible in the structure. The connections to the transmembrane regions of the chains are down in the drawing.

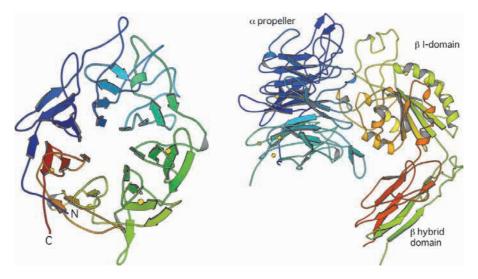


Fig. 16.6 • Left: The β propeller domain of the α chain of $\alpha V \beta 3$ integrin. The yellow spheres show the positions of four calcium ions bound at equivalent positions in four of the seven propeller blades. Right: The head portion of the ectodomain. The I domain of the β chain interacts with the β propeller of the α chain (PDB: 1JV2).

domain and the I domain regulates the activity of the integrin similarly to the way in which the propeller domain of the trimeric G-proteins blocks the activity of the $G\alpha$ subunit.

The stalk region consists of a number of Ig-type domains in the α subunit, called calf-1 and calf-2, and a number of EGF modules in the β chain. The C-terminal domain of the β chain has a unique fold and is often called the β -tail domain.

One large group of integrins has an insertion in the propeller domain of the α chain, which forms a second I domain with the same fold as the I domain of the β chain. This is mainly true for the α chains associated with β 2 (Table 16.2). In the integrins containing an α I domain, this domain is responsible for direct binding of ligands.

16.2.1.1 Metal binding motifs

The I domains of integrins bind to all their ligands in a unique way. The interaction is regulated by metal ion sites, one called the MIDAS (metal ion dependent adhesion site). Proximal to this site are two other metal binding sites, ADMIDAS (adjacent to MIDAS) and LIMBS (ligand associated metal binding site; or SyMBS). While calcium but not magnesium binds to the LIMBS and ADMIDAS sites, the MIDAS site prefers manganese or magnesium. Metal binding at the LIMBS stabilizes and is synergistic with binding at the MIDAS site. The MIDAS metal ion is coordinated by two serine and one aspartate residues (Figure 16.7). Metal ions like Mg²⁺ coordinate six oxygen atoms including two negative charges (Section 3.3.4.2). Since the arrangement of the oxygen ligands from the I domain at the metal ion is incomplete, another protein can bind to the metal ion by an additional negatively charged residue (Asp or Glu). Water molecules can otherwise complete the metal coordination.

The structure of a complex between an I domain and a collagen-like structure illustrates that the interaction with the metal ion through an acidic residue is an important part of the binding. The collagen-like triple helix binds to the MIDAS motif of the integrin through a glutamate side chain. The contact surface is large and shows both polar and non-polar interactions (Figure 16.8).

16.2.2 Mechanisms to Control Integrin Binding

How are the signals transmitted from the cytoplasm, through the transmembrane helices to the ectodomain or in the opposite direction? In one case it is the outside-in direction, where ligand binding influences the cytoskeleton inside cells, and in the other case it is the inside-out, where the status of the cytoplasm influences the binding affinity by the ectodomain.

The extracellular part of integrin can assume at least three different conformations. The bent conformation (inactive) and two extended forms, called closed and open (activated) have been observed. The closed and open forms differ in the relative position

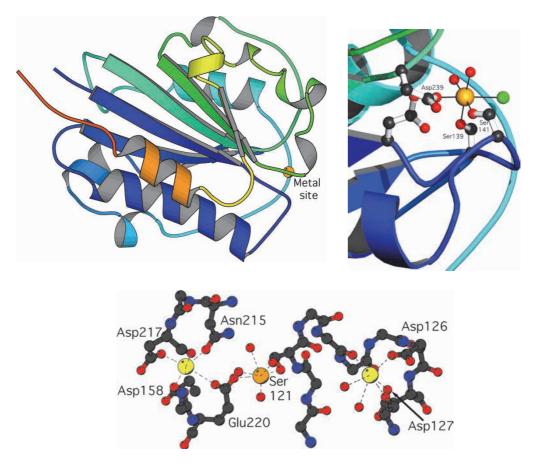


Fig. 16.7 ■ Left: Schematic drawing of an I domain from αL. In this case it comes from an insert into the propeller domain of αL (Table 16.2) Right: Close-up view of the metal binding. The metal ion (orange, a manganese ion in the crystal structure although magnesium ions are used for this function in vivo) is coordinated by two serine (below) and one aspartic acid (left) residue. In the crystal structure, two water molecules (red) and a chloride ion (green) are also bound, adding up to a total of six ligands. This is the closed conformation of the MIDAS motif (PDB: 1LFA). Below: The three metal binding sites from *left* to *right* LIMBS, MIDAS and ADMIDAS with Ca²⁺ in yellow and Mg²⁺ in orange.

of their transmembrane part. This conducts a signal into the cytoplasm. One of the proteins in the cytoplasm that plays a major role in the integrin signaling is talin. The binding of domain F3 of talin to the intracellular domain of the β subunit of integrin activates it for ligand binding. Conversely, ligands binding to the ectodomain will induce talin to bind in the cytoplasm. The most important difference to the inactive form is the change in angle between the BI domain and the hybrid domain (Figure 16.10). In the inactive form, the chains are close together.

The main difference between the low affinity state (closed) and the high affinity state (open) is the C-terminal helix (α 7) of the I domain, which is shifted by 9 Å



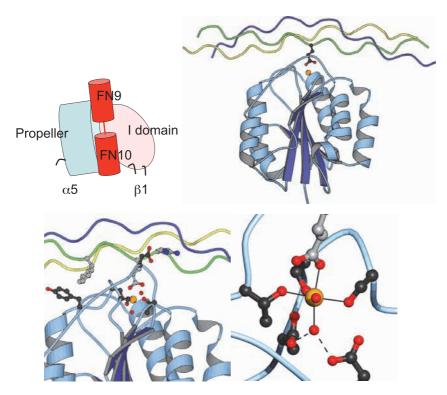


Fig. 16.8 ■ Binding of fibronectin and collagen to integrin. *Top left*: A model of how the headpiece of integrin α5β1 binds to fibronectin type III domains 9 and 10. Top right: The triple helix of collagen bound to the I-domain of integrin α2β1 (PDB: 1DZI). Bottom left: Close-up showing the position of the metal (orange; in this crystal structure a cobalt ion) and its oxygen ligands. Some side chains involved in direct contact between the molecules are included. The side chains from collagen have grey carbon atoms. Bottom right: Detail of the ligands at the metal. Oxygen atoms from two serine and a threonine residues from the integrin bind the metal ion (full lines). Two water molecules and a glutamate from collagen (grey carbon atoms) complete the coordination. Two aspartates from the integrin domain form hydrogen bonds (dashed lines) to one of the water molecules. This is the open, high-affinity conformation of the MIDAS motif.

(Figures 16.9 and 16.11). This movement is coupled to a change in the coordination of the metal ion in the MIDAS motif. In the closed state, the metal ion is coordinated by a negatively charged side chain and two uncharged side chains. In the open state, only uncharged residues bind directly to the metal ion, which leads to a higher affinity for the negatively charged binding partner (Figure 16.11).

The coupling of the shift of the α 7 helix and the metal coordination allows integrins to convert ligand binding into conformational change. The movement of the helix changes the relative orientation of the I domain and the hybrid domain. This corresponds to a shift of about 70 Å at the end of the PSI domain (Figure 16.10).

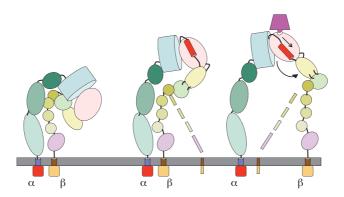


Fig. 16.9 ■ A schematic view of the possible conformation of integrin heads. The bent (*left*) and closed (middle) states are low affinity states, while the open state (right) is a high affinity state. The dashed lines illustrate that the complex is flexible. The C-terminal α -helix of the I domain (red) moves when going to the high affinity state and opens a binding site for a ligand (purple).

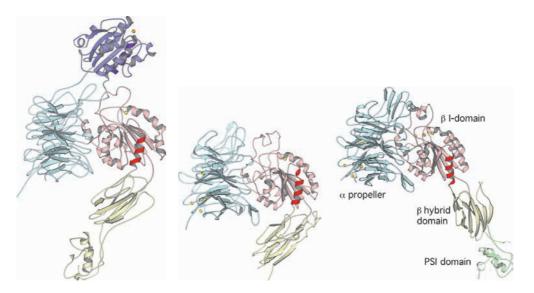


Fig. 16.10 • Left: The head portion of the $\alpha X\beta 2$ integrin, showing an α chain with an inserted I domain (PDB: 3K6S). The I domain inserted in the propeller domain (blue) with its C-terminal helix (purple, partly hidden). Middle and right: The heads of the αVβ3 (middle, PDB: 1JV2) and αΙΙbβ3 (right, PDB: 2VDK) integrins in the same orientation. The angle between the I domain and the hybrid domain is different and controlled by the C-terminal helix of the I domain (red).

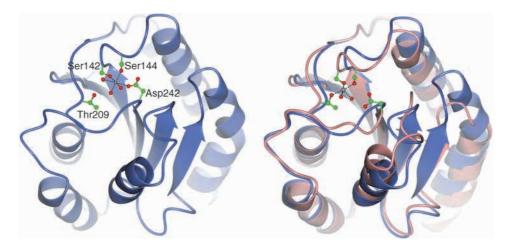


Fig. 16.11 ■ Left: The metal (gray) of the closed state in the I domain is coordinated by two serine and one aspartic acid residue, and three water molecules. Right: The coordination in the open activated state (pink) involves only uncharged residues; Thr209 has replaced Asp242. The closed state (blue) is superimposed to illustrate the conformational differences linked to the C-terminal (rightmost) helix (PDB: 1JLM and 1IDO).

In integrins with two I domains, it is the I domain in the α chain that is responsible for ligand binding. In these integrins, the coupling between ligand affinity in the αI domain and conformational changes in the β chain is indirect, and apparently involves a negatively charged side chain in αI that binds to the MIDAS motif in βI (Figure 16.10). If this binding controls the affinity of the binding site in αI , the intracellular status and the general conformation of the integrin molecule can regulate the activity in these integrins as well.

Recommended Reading

Original Articles

Dong X, Mi L-Z, Zhu J, et al. (2012). $\alpha_V b_3$ integrin crystal structures and their functional implications. Biochemistry 51: 8814–8828.

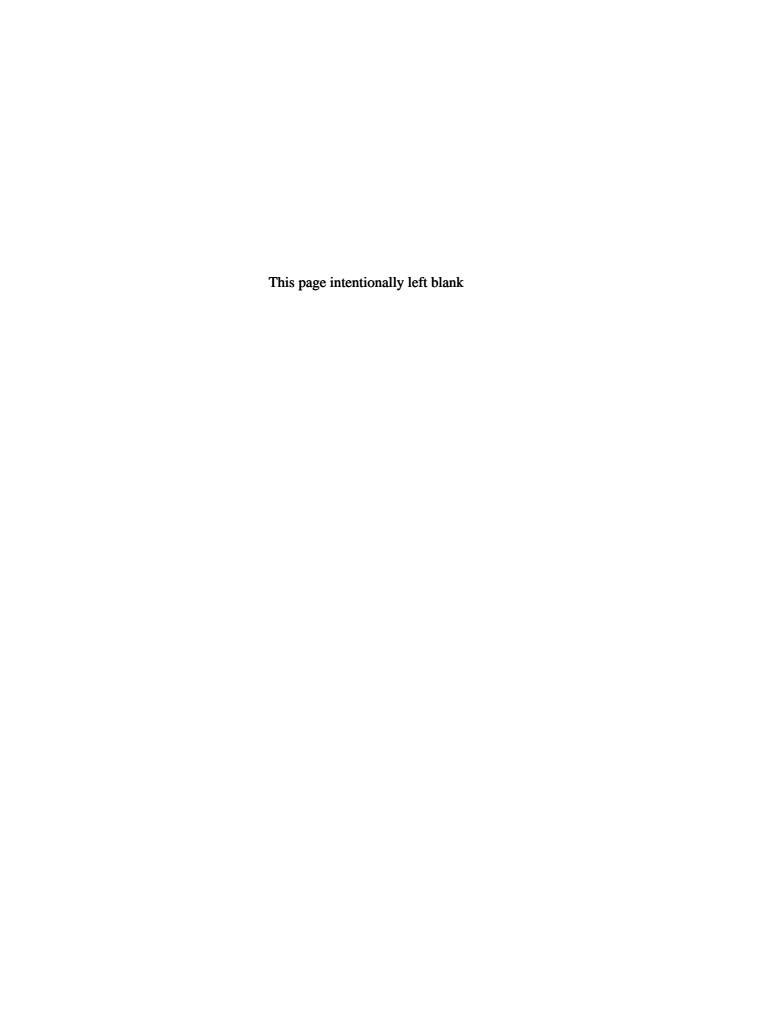
Emsley J, Knight CG, Farndale RW, et al. (2000) Structural basis of collagen recognition by integrin $\alpha_2\beta_1$. Cell **101**: 47–56.

Erat MC, Slatter DA, Lowe ED, et al. (2009) Identification and structural analysis of type I collagen sites in complex with fibronectin fragments. Proc Natl Acad Sci USA 106: 4195-4200.

- Nagae M, Re S, Mihara E, et al. (2012) Crystal structure of α5β1 integrin ectodomain: Atomic details of the fibronectin receptor. J Cell Biol 197: 131–140.
- van Agthoven JF, Xiong J-P, Alonso JL, et al. (2014) Structural basis for pure antagonism of integrin $\alpha_V \beta 3$ by a high-affinity form of fibronectin. *Nature Struct Mol Biol* **21**. 383–388.

Reviews

- Campbell ID, Humpfries MJ. (2011) Integrin structure, activation and interactions. Cold Spring Harb Perspect Biol 3: a004994.
- Das M, Ithychanda SS, Qin J, Plow EF. (2014) Mechanisms of talin-dependent integrin signaling and crosstalk. Biochim Biophys Acta 1838: 579-588.
- Luo BH, Carman CV, Springer TA. (2007) Structural basis of integrin regulation and signaling. Ann Rev Immunol 25: 619-647.



The Immune System

All organisms must defend themselves against infections by viruses or microorganisms. Cells also need to be able to distinguish self from non-self. These are complicated tasks, which require highly elaborate systems. All eukaryotes have a non-adaptive or innate immunity system, which is non-specific. Primary anatomical barriers are the skin and stomach, mouth and nose, the respiratory airways and lungs, and eyes. In addition, there is a large range of cellular and molecular defense mechanisms that will not be discussed here.

Vertebrates also have an adaptive immune system. The adaptive system defends the organism through two routes: the antibody-mediated system (humoral immunity) and the T-cell mediated system (cellular immunity). Adaptive immunity has two key properties. It is very specific to properly recognize foreign molecules and destroy parasites. In addition, it has a memory system helping the organism to respond rapidly if a reinfection occurs. Antibodies are produced by B cells and recognize surface epitopes on antigens. This humoral immunity deals with viruses and bacterial infections outside the cells. By contrast, cellular immunity deals with cells infected by viruses, bacteria or any type of foreign material or tissue. MHC molecules (Major Histocompatibility Complex) on the cellular surface present peptide fragments of foreign proteins to the T-cell receptors. Cells containing foreign material will be destroyed. The MHC molecules are also called transplantation antigens, which are so diverse that two individuals are unlikely to have the same set of MHC molecules. Therefore, transplanted cells are identified by the host as foreign due to the differences in their MHC molecules. Numerous structures of the proteins of the immune system have given us detailed views of the interplay between the central molecules.

17.1 Humoral Immunity — the Antibody-Mediated System

In humoral immunity, the specific recognition of an antigen by an antibody leads to the neutralization and destruction of the foreign substance or cell by phagocytosis or activation

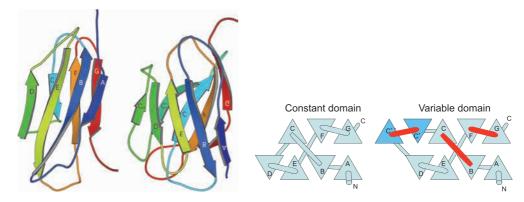


Fig. 17.1 ■ The immunoglobulin (Ig) fold. Many proteins in the immune system have this fold, as well as proteins involved in cell adhesion and the nervous system. Left: Ribbon representations of the fold of a constant domain and a variable domain (PDB: 1AQK, heavy chain). Right: A simplified representation of the β-sandwich that constitutes the Ig fold. The constant domain has a fourstranded and a three-stranded antiparallel sheet, but in the variable domain there are two extra β -strands, C' and C'' (darker blue). The red connections between some strands in the variable domain are the complementarity-determining regions, CDR1, CDR2 and CDR3, consecutively along the polypeptide chain. These regions form the antigen-binding surface.

of the complement system. Phagocytosis is the ingestion of the foreign material into white blood cells. The complement system consists of a sequence of serine proteases that activate one another. This finally leads to the killing of the foreign cell by lysis of its membranes.

To be able to specifically recognize the vast number of molecules, the antibodies or immunoglobulins produced by B cells have an enormous variability. How is this achieved? To explain this we need to describe the organization of the antibodies. Antibodies are built of repeats of the same type of domain, a β-sandwich with two layers of antiparallel β -strands (Figure 17.1).

17.1.1 IgG Molecule

Antibodies are built of two types of polypeptide chains, heavy and light. The light chains are always composed of two domains, each with about 110 amino acid residues. There are two main types of light chains, κ and λ . The heavy chains have at least four domains. In mammals, there are five different antibody classes with different functional properties and locations in the organism: IgA, IgD, IgE, IgG and IgM. Other vertebrates have a more limited setup of antibodies. They all have different types of heavy chains and can form different oligomers. In blood plasma, the IgG molecule is the most common type of immunoglobulin. It has two heavy and two light chains with a total of 12 Ig-domains (Figure 17.2). The amino terminal domains of a heavy and light chain pair form the antigen-binding domains. An IgG molecule has two identical such pairs which can bind antigens.

The Ig-domains normally interact in a pairwise manner. They can be hetero-pairs as in the pairs of domains between the heavy and light chains in the Fab fragments, or homo-pairs as in the Fc fragment (Figures 17.2 and 17.3). There are also homodimers of

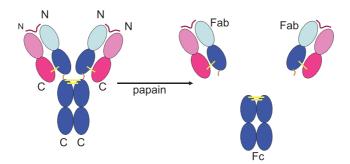


Fig. 17.2 ■ The IgG molecule is built of four multidomain chains, two heavy (blue) and two light chains (red). The heavy chains are composed of four domains, while the light chains have two domains. The light blue and red domains are the variable domains where the antigen binding occurs (purple). The darker domains are the constant domains. The heavy chains are linked to each other and the light chains are each linked to one of the heavy chains by disulfide bonds (yellow). If a proteolytic enzyme like papain treats IgG, the heavy chains are cleaved between the first and second constant domains such that three fragments are generated. The fragment composed only of heavy chain constant domains is called Fc (as in "fragment crystalline"). When it was first produced, it crystallized spontaneously in the dialysis tube!). The two identical fragments are called Fabs (fragment antigen-binding).

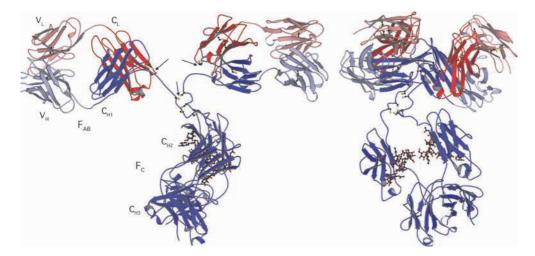


Fig. 17.3 • A detailed structure of IgG in two orientations. The heavy chain consists of the V_H , C_{H1} , C_{H2} and C_{H3} domains and the light chain of the V_L and the C_L domains. The Fab units are seen above and have very flexible links to the Fc unit below. The chains are connected by disulfide bonds. All cysteines that form disulfide bonds are shown and arrows indicate the ones that connect the chains. The Fc unit has carbohydrate modifications at the C_{H2} domain, drawn as ball-and-stick models (PDB: 1IGT).

light chains called Bence-Jones proteins produced in large quantities by certain types of cancer cells.

The antigen binding domains at the N-termini of the four chains are also called the variable domains (V_H or V_L). The C-terminal domains are called constant domains (C_H or C_L). While the constant domains normally are composed of two-layered β -sandwiches with

seven β -strands, the variable domains have two extra strands added to the layer with only three strands (Figure 17.1).

17.1.2 Recognition of Antigens

Antibodies are able to specifically recognize a very large numbers of antigens, of the order of millions. In a few days to a few weeks a vertebrate can produce large amounts of an antibody against a new foreign molecule. The complementarity determining regions (CDR1, CDR2 and CDR3, Figure 17.1) of heavy and light chains are responsible for the enormous capacity to recognize and bind different antigens. These six hypervariable loops of each antigen-binding surface form a contiguous area of around $1000 \, \text{Å}^2$, with flat parts as well as grooves and crevices for the specific binding of antigens.

Antibodies can bind small molecules called haptens. The binding of haptens, peptides and proteins has been characterized in terms of structures. The contact surfaces between antibodies and protein antigens were first analyzed in complexes between lysozyme or influenza virus neuraminidase with monoclonal antibodies against these proteins (Figure 17.4). The contact surface of the antibody involves most or all of the CDRs, which form a relatively flat area that is complementary in shape and polarity to the contacted area of the antigen. The conformational changes when the complex is formed are small.

17.1.3 Neutralizing Antibodies

Neutralizing antibodies (nAbs) defend a cell from an antigen or infectious bodies by preventing any damaging effects they may have. One example could be diphtheria toxin, which can be neutralized by a nAb. The nAb binds to the antigen and neutralizes its biological activity. Binding antibodies on the other hand signal to a white blood cell that the antigen has been targeted and needs to be destroyed.

In work with the virus HIV (Chapter 18), a new class of antibodies was identified, broadly neutralizing antibodies (bnAbs). The virus has a highly variable outer layer with

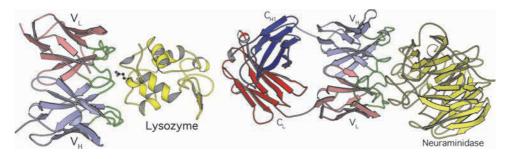


Fig. 17.4 ■ The binding of monoclonal antibodies to their antigens. *Left*: The variable domains of the Fab fragment bind a lysozyme molecule (PDB: 1VFB). *Right*: A Fab fragment binds the protein neuraminidase from the influenza virus (PDB: 1NCA). The hypervariable loops (CDRs) are shown in green.

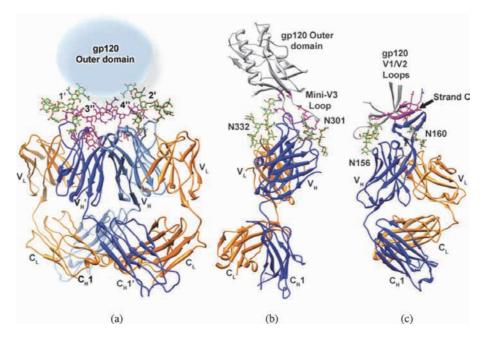


Fig. 17.5 ■ (a) The binding to the heavily glycosylated outer surface protein Env (gp120+gp41) of HIV interacting with broadly neutralizing antibodies (bnAbs) showing interactions with both carbohydrate and protein (PDB: 1OP5, 3TYG and 3U4E). In (b) and (c) the Asn residues to which the carbohydrate is attached is shown (N332, N301, N156 and N160). Longer hypervariable loops enable the antibody to get into contact with gp120. (Reproduced with permission from Julien J-P, Lee PS, Wilson IA. (2012) Structural insights into key sites of vulnerability on HIV Env and influenza HA. *Immunol Rev* **250**: 180–198. Copyright John Wiley & Sons A/S.)

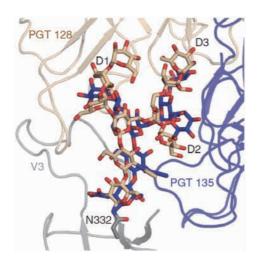
the protein Env, which is cleaved to gp120 and gp41. Env mediates membrane fusion with the host cells and is heavily glycosylated (Chapter 7). The variable loops of Env and the carbohydrates make it difficult for antibodies to bind to specific epitopes or antigenic determinants. However, the bnAbs have stickier and longer hypervariable loops, which pass the outer unspecific layer and identify viral epitopes. In work on gp120, it was observed that bnAbs can make specific interactions with the carbohydrate or combine the binding to both protein and carbohydrate (Figure 17.5).

The Ab interactions with carbohydrates depend on structural stability, as in the case of Asn332 in gp120. Several of the bnAbs studied bind at different positions around the GlcNAc₂Man₆ bound to N332 (Figure 17.6). This suggests that vaccines could be developed with antigens composed of glycosylated peptides.

17.1.4 Genomic Segments give Antibody Diversity

How can genomes like the human, with about 30 000 genes, generate millions of specific antibodies? The extensive antibody diversity is due to gene rearrangements. The human genome contains many gene segments that can recombine in different ways to become an

antibody. The light chains (κ or λ) are constructed from any combination of three different genetic elements, the V (variable) segment (more than 70 each for κ and λ), the J (joining) segment (five for κ and at least seven for λ) and the C (constant) segment, which is only present in one copy for κ but at least 7 for λ (Figure 17.7). For the heavy chains there are more than 100 V, 9 J and 11 C segments, and 27 copies of a D (diversity) segment that is



Fig, 17.6 ■ The glycosylated Asn332 of gp120 with protruding parts of bnAbs (PGT128 and PGT135) bound on either side of the carbohydrate moiety. The antibody binding is highly dependent on the carbohydrate. (Reproduced with permission from Kong *et al.* (2013) Supersite of immune vulnerability on the glycosylated face of HIV-1 envelope glycoprotein gp120. *Nat Struct Mol Biol* **20**: 796–803.)

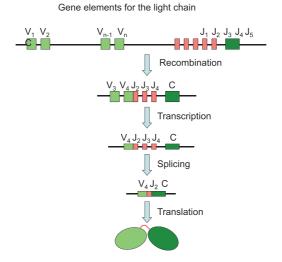


Fig. 17.7 ■ Light chain genetic rearrangement and expression. The different genetic segments can be combined in very many different ways. The undifferentiated cell has the complete set of antibody genes. During differentiation, a V gene becomes linked to a J gene in a random way. In the subsequent transcription of the DNA to a pre-mRNA, a further elimination is made. The mRNA is subsequently spliced into a mature mRNA that is translated into a specific light chain.

not found in the light chain. However, some of these segments are rarely, if ever used. The J segment in light chains or J and D segments in heavy chains code for part of CDR3. This increases the hypervariable character of CDR3, in particular for the heavy chains. In addition, the genetic linking is not always done in exactly the same place, which further increases the variation of the hypervariable loops. On top of this, frequent mutations in the CDR regions occur when B cells mature. The variation in heavy and light chains thus becomes very large and is further expanded by the random pairing of light chains with heavy chains.

During differentiation of immature B cells, IgM is first expressed and exposed on the cell surface. A single cell presents many copies of identical IgMs. Each cell is then targeted to a specific antigen. If a specific cell encounters an antigen, its IgMs will oligomerize. Through specific phosphorylation signaling and interactions with T cells, cell growth will be stimulated and further differentiation occurs. A large fraction of the B cells can then further differentiate and form plasma cells that secrete large numbers of antibodies specific to the antigen.

A vertebrate that has survived an infection by a certain pathogen is immune against further infections by the same pathogen. Normal T and B cells are short-lived but specific memory T and B cells will rapidly produce new cells carrying the immune defense against the pathogen if a new infection occurs.

17.2 Cellular Immunity — T-cell Mediated System

The T cells are one type of lymphocytes or white blood cells and are central for the cell-mediated immunity. The T-cell system involves the interaction between antigen-presenting cells and T cells. The antigen-presenting cells have major histocompatibility complex (MHC) proteins on the cellular surface. In humans, the complex is called human leucocyte antigen (HLA). The MHC molecules are divided into two classes, I and II. Both are found on antigen-presenting cells and display antigenic peptides on the surface of the cell (Figure 17.8).

Peptides bound to MHC I are cytosolic, while MHC II binds peptides from extracellular proteins. Since MHC I displays peptides of proteins from inside the cell on its surface and the T cells can identify whether these cells contain unsuitable proteins such as viral proteins or proteins from malignantly transformed cells, these should then be killed.

Peptides displayed by MHC II originate from outside the cell e.g. from bacterial or fungal infections. The foreign proteins can be endocytosed into the cell and fragmented in the lysosome and subsequently loaded on MHC II molecules to be presented on the cellular surface. When T cells identify the bacterial infection they trigger an appropriate

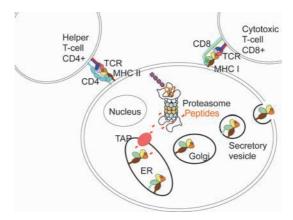


Fig. 17.8 ■ A schematic illustration of parts of the cellular machinery producing peptides to be presented by MHC class I or II, leading to interactions with different T cells. The MHC complexes are built up in the endoplasmic reticulum and transported through Golgi to the cell surface by secretory vesicles. Here the peptides they carry can be scrutinized by T-cell receptors. TAP is a membrane protein complex that transports the peptides generated by the proteasome into the ER to be loaded onto MHC I and II.

immune response. A range of molecular systems are involved in the transport, processing and loading of peptides onto the MHC molecules (Figure 17.8).

T cells have T-cell receptors and a range of additional molecules, primarily CD3 and CD4 or CD8, all of which connect to signaling systems. Structures of the central molecules are known, but their interactions are only partly understood.

17.2.1 MHC Presentation of Antigens

Both class I and II of the MHCs are heterodimers composed of four domains (Figure 17.9). Two domains have an Ig-type fold while the other two are $\alpha\beta$ domains that together form the site for peptide binding. In MHC class I, the peptide-binding site is associated with the α and α_2 domains of the heavy chain. The first structure determined was highly interesting. In addition to revealing the protein structure, a mixture of different peptides was bound to what was identified as the antigen-recognition site. The light chain, the β_2 -microglobulin subunit, is mainly associated with the α_3 domain of the heavy chain. In MHC class II, the two chains both contribute elements to the peptide-binding site (Figure 17.10). Nevertheless, the binding site for the peptide is designed in the same way in both types of MHCs. The base is built from an eight-stranded β -sheet with a helix on each side of the peptide. The two molecules most certainly must have a common origin in evolution. The molecular organization resembles a hot dog in a piece of bread.

MHC class I present peptides derived from intracellular degradation of proteins in the cytosol, whereas class II present peptides from degradation of extracellular antigens in endosomal compartments.



Fig. 17.9 ■ The structure of major histocompatibility complex (MHC) proteins of class I (*left*) and class II (*right*). MHC class I has a light chain, β2-microglobulin (β2m, green), and a heavy chain with three domains, α_1 (yellow), α_2 (orange) and α_3 (brown), where α_1 and α_2 form the binding site for the peptide. The third domain and β2-microglobulin both have the immunoglobulin constant region fold (PDB: 1A1M). MHC class II has a very similar arrangement of four domains in two chains, but the connections between the domains are different from those of class I (PDB: 1DLH).

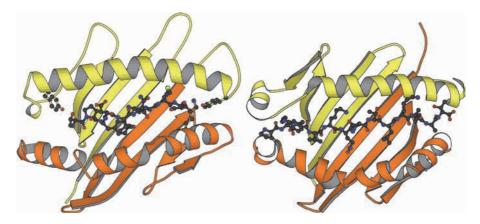


Fig. 17.10 ■ The binding of peptides to MHC class I (*left*) and class II (*right*) molecules. The peptide-binding site is a groove with a base of eight β -strands and two α -helices surrounding the peptide. The bound peptides are shown as a ball-and-stick figure. In MHC class I, some residues block the ends of the groove, while the ends of the groove are open in MHC class II.

Class I MHC molecules usually bind peptides 8–14 residues in length. The conformation of the peptide is extended with anchor residues bound in specificity pockets that differ in the alleles of MHC molecules. Since the ends of the binding site are closed, longer peptides will bulge when bound. In the class II binding site, the bound peptide adopts a

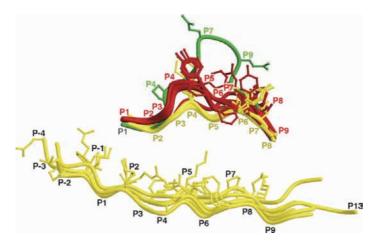


Fig. 17.11 ■ The structures of the peptides bound to MHC class I (above) and class II (below). The β -sheets of MHC have been aligned but are not shown. They are located below the peptides. The class I peptides are shown in different colors for different lengths: 8 (yellow), 9 (red) and 13 residues (green). The binding groove is closed at the ends in class I, therefore peptides of lengths longer than eight residues will bulge. For MHC class II, which lack the blocks at the end of the groove, the peptides can be longer and do not form any bulge. (Reprinted with permission from Rudolph MG, Stanfield RL, Wilson IA (2006). How TCRs bind MHCs, peptides and coreceptors. Ann Rev Immunol 24: 419–466. Copyright Annual Reviews.)

straight conformation, that of a left-handed polyproline helix. The binding site is open at both ends allowing larger peptides to protrude at either end. Thus, MHC class II can bind longer peptides than class I. There are also non-classical MHC molecules, which bind glycolipids and lipopeptides to be presented to T cells. The variation in the binding sites on different MHC molecules accommodates the wide range of peptides that needs to be presented (Figure 17.11). The side chains of some of the residues in the bound peptide are exposed and are accessible for interaction with T-cell receptors (TCR).

17.2.2 T-Cell Receptors

T-cell receptors (TCRs) are central in the immune surveillance in the body. They are located on the surface of T cells. Apart from a transmembrane region and a short cytoplasmic tail, they have the same general domain structure as antibody Fab fragments. They have constant and variable Ig-like domains and they are composed of α - and β -, or γ - and δ -chains. Disulfide bridges (in a manner similar to the Ig molecules) link both types. While the $\alpha\beta$ TCRs interact with antigenic peptides bound to MHCs, the $\gamma\delta$ TCRs bind directly to pathogen-derived glycoproteins or non-classical MHC molecules. As in antibody Fabs, the regions of the TCRs that interact with MHC plus bound peptide are called complementarity-determining regions (CDRs).

The CDRs interact with exposed side chains of the bound peptide, but also with the MHC α -helices that embed the peptide (Figure 17.12). The variable domain of the α chain $(V\alpha)$ is in contact with the N-terminal part of the antigen peptide, whereas V β contacts the C-terminal region (Figure 17.13).

17.2.2.1 CD8 or CD4 assist TCR in its interaction with MHC

TCRs are assisted in their interactions with MHC molecules by the coreceptors CD4 and CD8, which are also anchored in the T-cell membrane. CD8 is a heterodimer (the subunits are called α and β), where each monomer is composed of an Ig domain, a long linker and

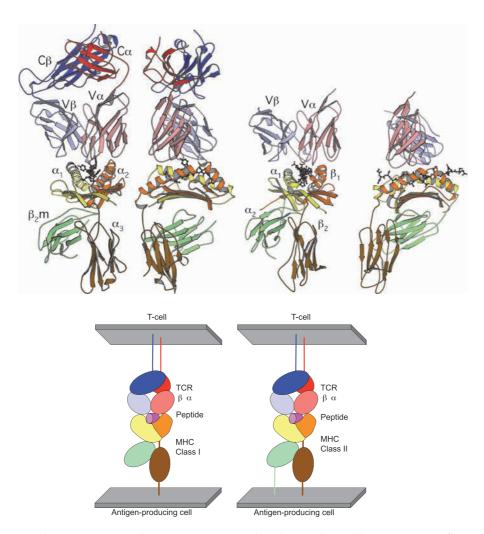


Fig. 17.12 ■ The interactions between MHC molecules and T-cell receptors. Left: Orthogonal views of the extracellular V and C domains of the T-cell receptor α and β chains bound to HLA-A201 (class I). A viral peptide is bound to the MHC molecule. Right: Variable domains of T-cell receptor D10 α and β chains bound to MHC I-A^k (class II). The T-cell receptor is seen above and the MHC molecule below with the bound peptide in the groove (PDB: 1BD2 and 1D9K). A schematic view of the interaction is shown below.

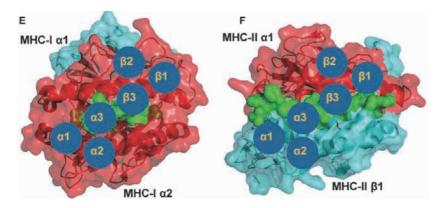


Fig. 17.13 ■ The interaction between MHC class I and II with TCR occurs at conserved sites (blue circles) on the surface of MHC class I and II. (Reproduced with permission from Holland CJ, Cole DK, Godkin A (2015) Re-directing CD4+ responses with the flanking residues of MHC class II-bound peptides: The core is not enough. Front Immunol 4 (172): 1–9.)

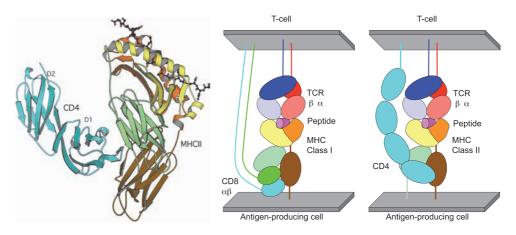


Fig. 17.14 ■ Left: The interaction between the D1 domain of CD4 (cyan) and the MHC class II complex. The bound peptide is shown as a ball-and-stick model (PDB: 1JL4). Right: Schematic view of how the TCR coreceptors CD8 and CD4 interact with MHC class I and II, respectively. The coreceptors bind to side of the MHC molecule, which is opposite to the peptide-binding site.

a transmembrane helix. CD4 is a monomeric protein composed of four Ig domains (D1-D4) of which D1 contacts MHC class II.

CD4 and CD8 interact with almost the same conserved regions on the underside of MHC class II and class I molecules, respectively (Figure 17.14). CD4 is also the primary cellular contact at infections with HIV1. CD4 interacts with the viral spike protein gp120. This interaction involves the same surface of CD4, but is much stronger than the interaction with MHC class II.

17.2.2.2 CD3 accessory molecules signal the state of the TCR molecules

There is an extensive range of proteins that are part of the signaling system related to the function of T cells. The TCRs have a very small intracellular domain insufficient for transfer of signals to the cellular machinery. Instead, TCRs are associated with three types of CD3 accessory molecules that contain domains involved in intracellular signaling. There are two types of heterodimers ($\gamma\epsilon$ and $\delta\epsilon$) of CD3, and these associate with the two chains of the TCR and a type of homodimer molecule ($\zeta\zeta$) into a complex of eight chains, each traversing the membrane. The extracellular domains of the CD3γε and CD3δε heterodimers consist of Ig folds interacting through joint β -sheets (Figure 17.15). The peptides that connect the Ig folds of the CD3s to the transmembrane region are relatively short and contain cysteines that, by forming disulfide bridges, assist in stabilizing the dimers. These short connections make the CD3 only reach the membrane-facing part of TCR. The transmembrane parts of CD3 contain conserved negatively charged residues that are important for the interactions with the positively charged transmembrane parts of the TCR molecules. The intracellular parts of CD3 contain short sequence motifs that are called ITAM (immunoreceptor tyrosinebased activation motif) and are involved in intracellular signaling. The tyrosyl residues can be phosphorylated by kinases of the Src family. The $\zeta\zeta$ has only nine residues on the extracellular side but have larger intracellular domains involved in signaling.

The antigen bound to MHC is identified by TCR and this information is subsequently transferred from TCR to the CD3 dimers through receptor aggregation and communicated further to the downstream signaling machinery. However, the communication between the extracellular and intracellular regions is only partly understood.

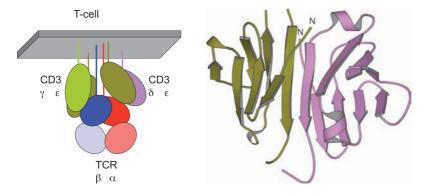


Fig. 17.15 • Left: A schematic illustration of the interactions between TCRs and CD3s in T cells. The CD3γε and CD3δε are heterodimers that interact with TCR. Their location in the membrane defines their interactions and the intracellular signals transmitted. Right: The extracellular domains of the CD3ε/δ dimer associate through a joint β-structure and with an approximate two-fold axis that is vertical in this view. The ε/γ dimer is formed in the same way. The N-termini leads to the transmembrane region (PDB: 1XIW).

Recommended Reading

Reviews

- Brazin KN, Mallis RJ, Das DK, et al. (2015) Structural features of the αβTCR mechanotransduction apparatus that promote pMHC discrimination. Frontiers Immunol 6: 1–13.
- Julien J-P, Lee PS and Wilson IA. (2012) Structural insights into key sites of vulnerability on HIV Env and influenza HA. *Immunol Rev* **250**: 180–198.
- Li Y, Yin Y, Mariuzza RA. (2013) Structural and biophysical insights into the role of CD4 and CD8 in T cell activation. Frontiers Immunol 4: 1–11.
- Mayerhofer PU, Tampe R. (2015) Antigen translocation machineries in adaptive immunity and viral immune evasion. J Mol Biol 427: 1102–1118.
- Rudolph MG, Stanfield RL, Wilson IA. (2006) How TCRs bind MHCs, peptides and coreceptors. Ann Rev Immunol 24: 419-466.

Original Articles

- Kong L, Lee JH, Doores KJ, et al. (2013) Supersite of immune vulnerability on the glycosylated face of HIV-1 envelope glycoprotein gp120. Nat Struct Mol Biol 20: 796–803.
- Stern LJ, Brown JH, Jardetzky TS, et al. (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 368: 215–221.
- Garcia KC, Degano M, Stanfield RL, et al. (1996) An αβ T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science* **274**: 209–219.

Virus Structure and Function

18.1 Virus Composition

Viruses are entities with genetic material enclosed in a protective shell of proteins. Some viruses are more complex in that they also have a membrane layer enveloping the nucleic acid (enveloped viruses). They depend on a living cell for synthesis of new proteins and are normally specific for one kind of host, but there are viruses infecting all types of living cells: bacteria, archaea, fungi, plants and animals. Viruses are found with very different sizes and shapes (Table 18.1). Their genome can be either DNA or rRNA. These nucleic acids can be either single-stranded or double-stranded. The genome can be in one or several segments and can code for as many as several hundred proteins or as few as four or five. The structural proteins form the virus particles that infect new host cells, and the non-structural proteins are mostly produced only in the infected cell and used for efficient production of the components of the particles. For example, viruses with an rRNA genome always code for an enzyme that can catalyze the replication of their genetic material, since host cells are normally not capable of doing that.

18.1.1 Symmetry of the Protein Shell

The protective shell around the genome is formed by protein molecules (Figure 18.1). In enveloped viruses, the lipid membrane is taken from the host cell when these viruses leave the host through a process called budding. The membrane contains virally encoded proteins that form the outer surface of the virus particle. The mechanism for entry into the host cell depends on the type of protective coat: enveloped viruses enter by fusing its membrane to the cell membrane or the membrane of an organelle, but non-enveloped viruses must use other mechanisms to enter cells.

TABLE 18.1 Composition of a Few Well-Known Human Viruses, Representing Different Sizes and Genome Types

| Name | Type of Genome ^a | Estimated Number of Proteins Coded for by Genome | Composition of Shell |
|--------------------|---------------------------------|--|--|
| Poliovirus | Single ss+RNA chain | Four structural, four non-structural proteins | Single icosahedral protein shell, 60 copies of three proteins |
| Influenza virus | Eight ss –RNA chains | Five structural, five non-structural proteins | Enveloped, outer surface formed by membrane proteins hemagglutinin and neuraminidase |
| Rotavirus | Eleven ds RNA molecules | Six structural, six non-structural proteins | Three layers of protein, with the inner two composed of 120 and 780 copies of coat proteins |
| HIV | Single ss+RNA chain | Five structural, 10 non-structural proteins | Enveloped, outer surface formed by glycoproteins SU and TM |
| Variola (smallpox) | Single linear dsDNA molecule | About 200 proteins | Very large, enveloped |

^aRNA genomes are single-stranded (ss) or double-stranded (ds) and are either positive-stranded (+), or negative-stranded (-). The RNA has to be replicated before protein synthesis. There are also human viruses with ssDNA (parvoviruses). Even larger viruses are found in the Mimiviridae family infecting amoeba. Their genomes code for about 1000 proteins, more than some types of bacteria.

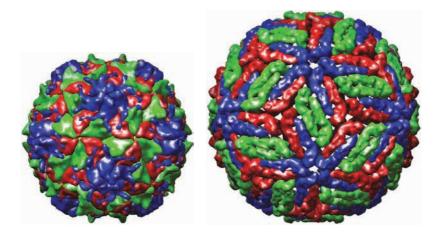


Fig. 18.1 ■ Schematic drawings of virus particles. *Left*: Poliovirus, a simple icosahedral virus with a diameter of about 300 Å (based on a crystal structure). Right: Flavivirus, an enveloped virus with a crystal diameter of about 470 Å (based on a cryo-EM model with models of coat protein molecules from a crystal structure fitted into the cryo-EM density). The colors denote subunits in different environments as discussed below. From VIPER (http://viperdb.scripps.edu/).

A nucleic acid molecule cannot code for a protein large enough to cover it, and this makes it necessary to form the protein shell using multiples of identical protein molecules. The need to protect the viral genome requires a stable outer shell. To achieve this, the protein subunits of the shell have symmetric arrangements, where the same protein-protein contact surfaces are used. There are essentially two kinds of symmetry found in viruses, helical symmetry and icosahedral symmetry. Helical symmetry leads to rod-shaped virus particles as in the tobacco mosaic virus and icosahedral symmetry leads to closed shells with a more or less spherical shape.

The shape of the virus particle depends on the composition of the shell. Some enveloped viruses have a variable shape due to the flexibility of the membrane, while others have a defined shape based on repeated interactions between the membrane proteins themselves or with an inner symmetric protein layer. Non-enveloped viruses normally have either helical or icosahedral symmetry. Some have a more complex shape. One group of viruses, the large DNA phages, have icosahedral or elongated heads and helical tails. In phage T4, these parts are complemented with other protein complexes forming fibers and other structures important in the infection process (Section 18.4).

18.1.2 Quasi-Equivalence

The icosahedron is an object formed by 20 equilateral triangles, and it has five-fold, three-fold and two-fold symmetry. Sixty identical units with the same environment are needed to generate an icosahedron (Figure 18.2). This is the highest possible symmetry in a closed object. (The dodecahedron formed by 12 pentagons has the same type of symmetry.)

Some virus shells are formed by 60 copies of a coat protein, but most virus particles are formed by larger numbers of identical subunits. In 1962, long before detailed structures of viruses were known, Caspar and Klug presented the quasi-equivalence theory, which tried to explain how large numbers of coat proteins could be arranged with

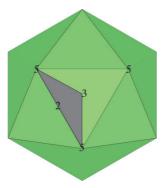


Fig. 18.2 ■ An icosahedron showing the positions of the 5-, 3- and 2-fold symmetry axes. The repeated unit is marked in grey. This is only one of many possible choices of the repeated unit.

icosahedral symmetry. If multiples of 60 subunits are packed in a shell, identical protein molecules will have different environments. The theory is based on the assumption that the contacts between protein subunits are similar, quasi-equivalent, and use the same bonds with slight deformations.

The basis of the theory is that it is possible to form six- and five-fold interactions with similar contacts between subunits. A plane triangular net with six-fold contacts can be transformed into an icosahedron if some of the six-fold contacts are replaced by five-fold contacts in a regular manner. The five-fold contacts create curvature, and depending on the position of these five-fold axes, icosahedra with different numbers of triangles are formed. Caspar and Klug found that certain multiples of 60 subunits could be arranged in this way and thereby maintain icosahedral symmetry. These multiples correspond to the triangulation numbers T = 1, 3, 4, 7, 9, 12, 13, 16, 19, 21, 25 and so on, following the scheme $T = h^2 + k^2 + hk$, where h and k are integers (Figure 18.3). The total number of subunits in the shell is 60T.

The predictions of the quasi-equivalence theory regarding the arrangement of protein subunits have mostly been confirmed by structural studies. Crystal structures of virus particles with triangulation numbers 1, 3, 4, 7 and 13 have made it possible to analyze quasi-equivalence at the atomic level (Figure 18.4). In some cases, the subunit contacts are indeed quasi-equivalent in the sense predicted by Caspar and Klug. In other cases, the quasi-equivalent contacts are mostly formed between different sets of atoms even though the positions of the subunits on the surface of the virus follow the rules of the theory. There are also some cases where the deviations from the predicted similarity of contacts are still larger. In the polyomavirus and SV40, which have shells with T = 7 quasi-symmetry, the six-fold positions as well as the five-fold positions are occupied by pentamers of subunits. The particle thus has 360 protein subunits rather than the predicted 420. In the blue-tongue virus and related viruses with two layers of proteins, the inner layer has 120 copies of a coat protein. This would correspond to a triangulation number of T = 2, which is not allowed by the theory. For these viruses, the interactions of the two subunits are widely different from each other.

18.1.3 Controlling Particle Assembly and Stability

In many simple viruses, the particles form from their components without the involvement of other molecules. This self-assembly must therefore somehow be an inherent property of the components. The coat protein molecules of icosahedral viruses are mainly responsible for the shape and size of the virus particles and are able to form five-, three- and two-fold contacts. When multiples of 60 chemically identical subunits form the shell, the molecules must be able to form at least slightly different contacts in a correct way to make well-ordered capsids with icosahedral symmetry. The first structures of virus particles to be determined were plant viruses with T = 3 quasi-symmetry. The

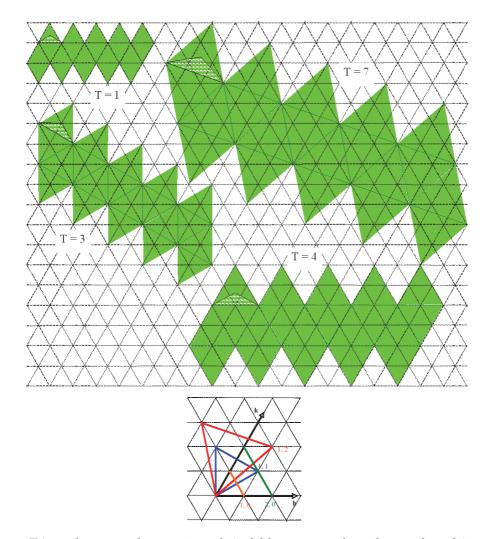


Fig. 18.3 ■ Triangular nets where points of six-fold symmetry have been selected in a regular manner to be replaced by five-folds. Twenty triangles (in green) correspond to the surfaces of the icosahedron, where each corner of the triangles corresponds to a five-fold axis. The asymmetric unit (as in Figure 18.2) is shown. The arrangements for triangulation numbers 1, 3, 4 and 7 are shown. The six-fold symmetry becomes a quasi-six-fold that coincides with the icosahedral threefold (T = 3) or two-fold (T = 4). In T = 1, only five-fold symmetry is found and in T = 7 the quasisix-fold does not coincide with any of the icosahedral symmetry axes. Bottom: The h-k- and coordinate systems in the net, showing one triangle for each of these triangulation numbers (orange: T = 1, blue: T = 3, green: T = 4, red: T = 7).



Fig. 18.4 ■ Viruses with triangulation numbers 1, 3, 4, 7, 13, and adenovirus (corresponding to T = 25, but with trimers of the hexon protein on the six-fold positions) showing their relative sizes. The surface of the virus particles is colored according to its distance from the center. All particles have icosahedral symmetry. The drawings are based on the crystal structures of (from left to right, top row) satellite tobacco necrosis virus and phage MS2 (bottom row) Nudaurelia capensis ω virus, phage HK97, blue-tongue virus and adenovirus. From VIPER (http://viperdb.scripps.edu/).

coat proteins had globular domains and an extended segment at the N-terminal end of the polypeptide chain.

The globular domain of the coat protein of these and most other non-enveloped icosahedral viruses has a fold formed by two antiparallel four-stranded β -sheets with the jellyroll topology (Figure 18.5, Section 3.2.3.1). The lengths of the strands and the length and conformation of the connecting loops vary widely between the different viruses. In some viruses, inserted domains interrupt the jellyroll fold.

In a T = 3 virus, chemically identical subunits will have three distinct environments. For example, one of the subunits forms five-fold contacts while the other two form quasi-six-fold contacts at the icosahedral three-fold axis. In the plant T = 3, viruses first studied, the extended segment of the coat protein molecules is used in the packing to stabilize the arrangement of the subunits. The N-terminal part of the coat protein is completely disordered in two of the three subunits, while it is partly ordered in one of them (Figure 18.6). These ordered arms are inserted in the interfaces between some of the subunits and interact with each other at the three-fold symmetry axis to form a structure that has been labeled the beta annulus. At the interfaces where the ordered arms are inserted, the details of the subunit-subunit contacts are very different from the quasiequivalent contacts formed by a disordered arm. The N-terminal arm acts like a switch guiding the packing, to achieve the correct curvature. The N-terminal arm and the order/disorder switching is important for determining the size of the particle. Mutants of the coat protein with the switching part of the arm removed only form T = 1, but no T = 3 particles.

Extended arms (N-terminal or C-terminal) and order/disorder switching have been found in most non-enveloped icosahedral viruses, in T = 3 as well as in T = 4 and T = 7

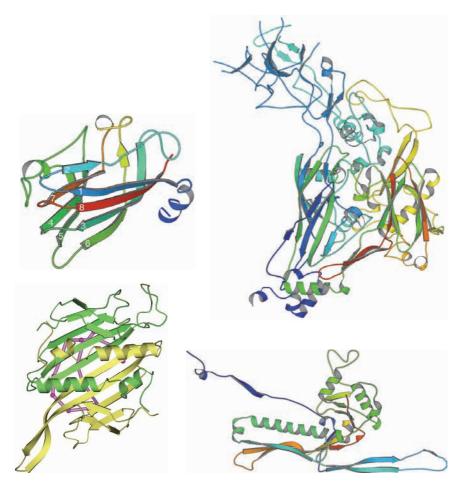


Fig. 18.5 ■ The most common coat protein folds found in icosahedral non-enveloped particles. Top left: The jellyroll fold in a viral coat protein subunit (satellite tobacco necrosis virus, PDB: 2BUK). The N-terminal helix (blue) extends into the interior of the particle, interacting with the RNA. Top right: A double jellyroll found in adenovirus and several large capsids from viruses infecting all kingdoms of life. The view is roughly tangential to the virus surface (adenovirus hexon protein, PDB: 3TG7). Bottom left: The MS2 fold, found in a family of small RNA bacteriophages. The view is radial (PDB: 2MS2). Bottom right: The HK97 fold, found in the capsids of tailed DNA bacteriophages (PDB:10HG).

viruses. The details of these switches are, however, quite different in different groups of viruses. There are also examples of viruses with quasi-equivalent arrangement of subunits where there are no extended arms. In these cases, the fold of the coat protein is not the common jellyroll. The extended arms might thus be characteristic of the jellyroll type of fold. The fact that the disordered arms are so common might be due to the common ancestry of these coat proteins rather than a functional necessity.

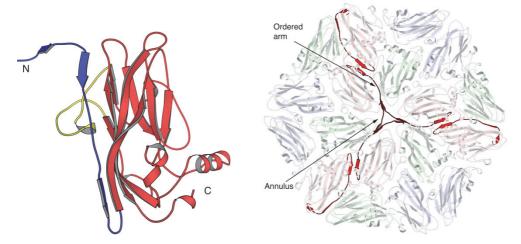


Fig. 18.6 ■ Packing of subunits in a T = 3 virus. *Left*: The jellyroll (red) and the N-terminal arm (blue) of the coat protein of southern cowpea mosaic virus (formerly called southern bean mosaic virus, PDB: 4SBV). The drawing shows the conformation of the C subunit where the arm is partly ordered. Right: The arrangement of 18 subunits around the three-fold (quasi-six-fold) axis in the southern cowpea mosaic virus. The partially ordered arm in one of the subunits (marked in red) interacts with arms from symmetry-related subunits at the three-fold axis (beta annulus, indicated with an arrow). In this virus, the N-terminal 23 amino acids are disordered in all subunits. This region contains several positively charged residues and probably interacts with the viral RNA, which is asymmetric.

Mechanisms for Host Cell Entry

18.2.1 Membrane Fusion

Viruses have evolved to enter their host cells in many ways. Once inside the host cell, they shut down the production of cellular components and gear the cellular machinery to produce new virus particles. Animal viruses need to pass the cell membrane, and for enveloped viruses this is achieved by membrane fusion: the membrane of the virus and the cellular membrane are joined, and the nucleocapsid containing the viral genome gets into the cell (see Section 6.6.3). Depending on the character of the viral genome, it is transported to the nucleus (DNA viruses) or remains in the cytoplasm (RNA viruses).

All of the events in the entry are programmed by the virus. The virus particles are relatively simple complexes of a few types of molecules, and the entry program depends on various kinds of triggers induced by interactions with the host cell, which lead to conformational changes in the particles.

Enveloped viruses use fusion peptides to join the viral lipid bilayer with the cellular or vesicular membrane. These peptides are part of viral surface proteins. There are several steps in the mechanisms leading to membrane fusion. Initially, the fusion peptide is hidden, but some kind of trigger exposes the peptide, often at the outer surface of an extended, trimeric structure. In this extended intermediate, the fusion peptide is bound to the cellular membrane and the protein is attached to the viral membrane through its C-terminal transmembrane helix. This extended structure changes conformation dramatically to bring the fusion peptide and the cellular membrane close to the viral membrane and fusion can occur. The fusion protein is thus able to undergo at least two steps of large conformational changes.

18.2.1.1 Class I fusion — controlled by proteolytic cleavage

There are three types of mechanisms controlling the availability of fusion peptides. In class I, represented by influenza virus and human immunodeficiency virus, HIV, the protein responsible for fusion is cleaved by a proteolytic enzyme. This is a priming step, leading to a metastable state. The mechanism for membrane fusion in influenza virus was the first one that was studied. Here, the fusion protein hemagglutinin is a glycosylated, membrane-bound, large protein that forms trimeric spikes on the viral surface. In the virus particle, the protein is cleaved at one specific position. The C-terminal part (HA2) anchors the protein to the viral membrane and forms one very long and one short helix, connected by a long loop (Figure 18.7). Part of the long helix forms a triple-helix coiled coil with the helices from the other subunits in the trimer. At the N-terminus of HA2, the fusion peptide is found hidden in the interior of the trimeric structure. The N-terminal part (HA1) forms the receptor domain and extends along HA2. Influenza virus binds to a receptor that leads to the formation of an endocytotic vesicle bringing the complete virus particle into the cell.

Induced by the low pH in endosomes, a very large conformational change occurs. A crystal structure of a low pH form has been determined. Part of the protein that forms a loop at neutral pH extends the long helix in the molecule, leading to exposure of the fusion peptide. At the other end, a segment of the helix is bent to become antiparallel to the long helix. Since this segment is connected to the membrane anchor, this jackknife movement is thought to bring the viral membrane close to the host cell membrane where the fusion peptide is inserted.

HIV particles may fuse directly with the cellular membrane, but the pathway used in in vivo infection may still involve endocytosis. The fusion protein is called env or gp160 and is cleaved by a protease before release from the host cell into two cleavage products, gp120 and gp41. The fragments are still associated after the cleavage. gp120 is exposed on the viral surface and interacts with the CD4 receptor on the target cell. The mechanism also requires binding to a co-receptor, the chemokine receptor CXCR4 (or CCR5). gp41 has a C-terminal transmembrane helix and participates in the fusion with the target cell. As in the case of influenza virus, large conformational changes in a trimeric complex of gp41 brings the fusion peptide, bound to the host membrane, and the transmembrane helix in the virus particle close in space.

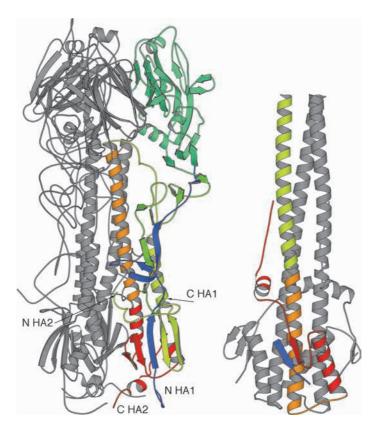


Figure 18.7 ■ Influenza virus hemagglutinin. The coloring is the same in the two drawings, and only one of the subunits is colored. Left: The trimer at neutral pH after cleavage into the HA1 and HA2 peptides (PDB: 3HGM). The membrane anchor if it were present would be attached to the C-terminus of HA2. The fusion peptide (N-terminus of HA2) is at the center of the molecule, hidden by interactions with the other proteins in the trimer. Right: The trimer of the low pH form (PDB: 1HTM). The fusion peptide is disordered, but must be located on top of the three helices. Only a very short segment of HA1 is present (blue). Large conformational changes have occurred. The lower part of the long helix in the neutral pH form (red) is now bent and packed at the side of the long helix. Note that a small part of the five-stranded β -sheet close to the membrane in the neutral form is retained in the low pH form (red and blue strands) but rotated together with the helix. The upper part of the long helix (green-yellow) forms a long loop and a helix in the neutral pH form. The visible C-terminal of HA2 (red) is now pointing up in the structure, in the same direction as the disorder fusion peptide at the N-terminal.

18.2.1.2 Class II fusion — controlled by a chaperone

Another type of fusion peptide is found in flaviviruses and alphaviruses, which are enveloped viruses with icosahedral symmetry. The fusion peptide is hidden in subunitsubunit interactions in this group of viruses as well. The elongated membrane protein,

called E in flaviviruses and E1 in alphaviruses, has a completely different conformation to the influenza virus hemagglutinin (Figure 18.8). The fusion peptide is a loop between two beta strands. In its non-activated form, the protein forms a dimer with a tangential orientation with the fusion peptide bound to one domain of the other subunit. This structure is

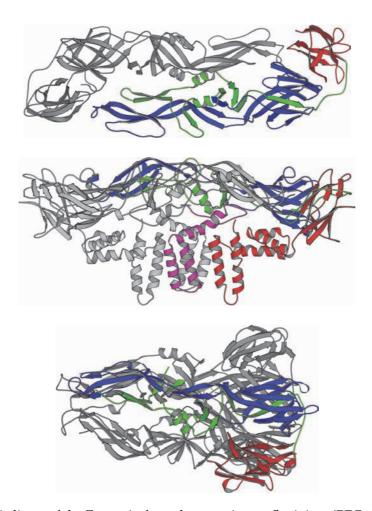


Fig. 18.8 ■ *Top*: A dimer of the E protein from dengue virus, a flavivirus (PDB: 1OAN). The complete layer of E proteins in the particle is shown in Figure 18.1. The coloring is from N-terminal (blue) to C-terminal (red). The fusion peptide is the loop at the extreme left of the colored molecule and is hidden through contacts to the other protein in the homodimer (grey). Middle: A dimer of the E protein in a view tangential to the membrane surface (PDB: 3J27). One subunit is colored as in the top view. This is a cryo-EM structure of the complete particle and includes also the membrane-bound helices. The anchor to the viral membrane is at the C-terminus of the protein (red helices in one of the subunits). The chaperone protein M is also shown (purple in one of the subunits). Bottom: The trimeric structure after fusion. The fusion peptide is the leftmost loop structure. The C-terminal domain (red) has rotated in relation to the rest of the protein (PDB: 3G7T).

18.2.1.3 Class III fusion

A third type of fusion mechanism is found, for example, in vesicular stomatitis virus and herpes simplex virus. These viruses have no proteolytic cleavage and no chaperone protein. The mechanism involves formation of a trimeric structure that brings the viral and host membrane in proximity, just as in the other classes of fusion.

18.2.2 Entry Mechanisms for Non-Enveloped Viruses

Non-enveloped viruses have to use methods other than fusion to penetrate a membrane in the host cell. One possibility is that the virus is able to form a pore in the host cell that allows the particle or the viral genome to pass through the cellular or endosomal membrane.

Picornaviruses, including poliovirus and rhinovirus (the common cold virus), are simple viruses with only a few structural and non-structural proteins. Their entry mechanisms have been studied as one of the possible ways of finding drugs to prevent viral infections. The mature poliovirus particle is built up of 60 copies of each of the three coat protein molecules VP1, VP2 and VP3. All three coat proteins have jellyroll folds and relatively long N-terminal extensions that form a network on the inside of the particle (Figure 18.9). A separate polypeptide, VP4, is found inside the protein shell. This peptide is initially an N-terminal extension of VP2, but is cleaved by autocatalysis after particle formation. This type of "maturation cleavage" is common in viruses. The virus particles provide an environment that favors the cleavage. They are not very efficient as catalysts, but there is no need for a fast turnover.

Infection starts with the virus particle binding to a receptor molecule on the surface of the host cell. For the poliovirus, the receptor is a cell adhesion protein called CD155 or PVR, poliovirus receptor. Interaction of the virus with the receptor appears to induce large conformational changes in the virus. The N-terminal tail of VP4 is myristoylated and will therefore have an affinity for membranes. Experiments have shown that after binding to the receptor, VP4 is lost from the particles and the N-terminal tail of VP1 is externalized. This shows that the particles are very dynamic, allowing holes to be created in the capsid, which leads to the release of the peptides from the interior. These changes might therefore be steps in the formation of a pore in the membrane through which the rRNA can enter the cytoplasm. The details of this entry process are, however, still unknown.

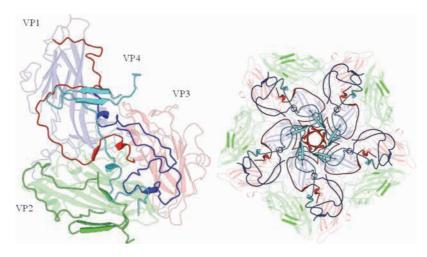


Fig. 18.9 ■ N-terminal arms in the poliovirus. *Left*: The repeating unit (protomer) as seen from the inside of the shell. The N-terminal extensions are shown in dark shading, while the main part of the subunit is pale. The N-termini of VP1 and VP3 are bound to the main parts of VP3 and VP1, respectively, while the remaining N-terminal of VP2, after cleavage of VP4, is bound to the subunit itself. Right: Packing of subunits around the five-fold axis as seen from the inside of the particle. VP1 forms the outside surface and the hole at the axis is "plugged" by the N-terminus of VP3 (red). Upon receptor binding, VP4 (turquoise) leaves the particle and the N-terminus of VP1 (blue) becomes exposed.

Binding to Nucleic Acids 18.3

In the assembly process, it is important that the correct nucleic acid molecule is incorporated into the infectious particle. Different mechanisms might be used for this, but one of the best-known cases is phage MS2. This virus is a small, simple T=3 virus composed of 180 copies of a coat protein and a single-stranded rRNA molecule of about 3500 bases. The coat protein acts as a translational repressor of one of the viral genes by binding to an rRNA hairpin of about 19 nucleotides, which includes the initiation codon of the replicase gene. This binding is also the initial step in particle assembly. The rRNA hairpin is thus both a translational operator and an encapsidation signal.

The MS2 coat protein has a unique fold with a single sheet and two helices. Two monomers form a dimer through very strong interactions, where the helices are inserted in a pocket in the other subunit (Figure 18.5).

Many rRNA-binding proteins are built up of an antiparallel sheet and a few helices. Although similar to them, the MS2 coat protein does not have the same topology as other RNA-binding proteins. It binds rRNA at the surface of the sheet (Figure 18.10). The binding of the rRNA is through single-stranded regions of the molecule, i.e. the four-nucleotide loop closing the hairpin and a bulged A nucleotide in the stem.

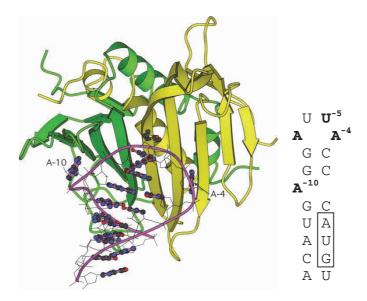


Fig. 18.10 ■ Left: The binding of the RNA hairpin by the MS2 dimer (PDB: 1ZDI). Adenine bases -10 and -4 are bound in corresponding pockets in the two monomers of the dimer, and uracil base -5 is stacked to a tyrosine sidechain in one of the subunits. Right: The secondary structure of the hairpin is shown. The initiation codon of the replicase subunit is boxed.

18.4 Structure of a Complex Virus: Phage T4

Bacterial viruses do not enter into their hosts but are able to deliver their genome to the inside of the bacterial cells using different mechanisms. The small rRNA phages described above insert their rRNA through pili that normally are used for exchange of genetic material between bacteria. A very large group of DNA phages have tails that are used to inject the viral DNA. The most complex family of tailed phages is the myoviruses, to which phage T4 belongs (Figure 18.11). This virus has a head that is a T = 13 icosahedron, where an extra ring of subunits makes it elongated. The tail ends with a baseplate to which two types of fibers are attached. The baseplate has partial six-fold symmetry and is formed by at least 14 different proteins, most of them present in several copies. The understanding of the tail structure and its function comes from a combination of electron microscopy, crystal structures of individual proteins or protein oligomers, and a vast number of biochemical experiments (Figures 18.12 and 18.13).

The assembly of the complete particle follows a specific pathway where each component is added in turn. The head is assembled using scaffolding proteins that are degraded and leave before a portal protein injects the DNA. The tail is assembled separately and joined to the DNA-packed head.

In the mature virion, the tail is loaded like a spring. Interactions between the short tail fibers and the bacterium lead to conformational changes in the baseplate. These changes

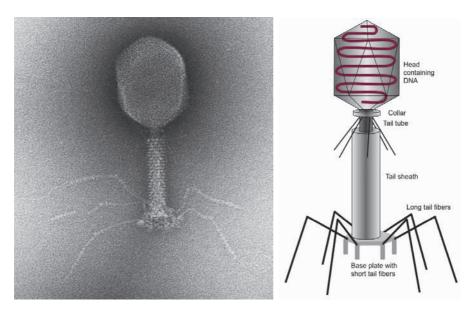


Fig. 18.11 ■ Phage T4. Left: An electron micrograph of the phage. (Courtesy of R. Duda, Pittsburgh.) Right: The main parts of the virion are labeled as in the schematic view.

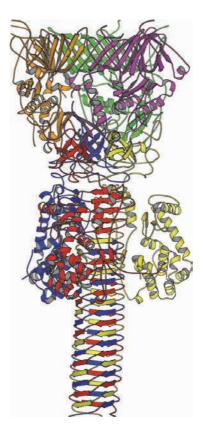


Fig. 18.12 ■ The trimeric gp5-gp27 complex. The three monomers of gp5 are in red, blue and yellow, and the monomers of gp27 in green, brown and purple. The lysozyme domain of gp5 is at the upper part of the triple beta helix that forms the stalk of the molecule.

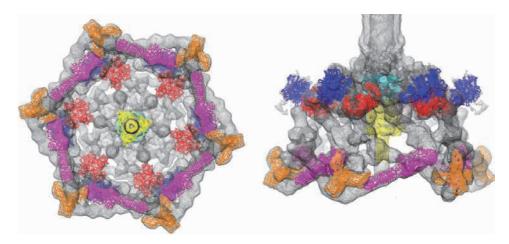


Fig. 18.13 ■ Fitting of several proteins from the T4 baseplate into a cryo-EM map. The gp5-gp27 complex is at the center of this model. gp5 is in yellow and gp27 (barely visible) in turquoise. The other proteins that are modeled are gp9 (blue), gp8 (red), gp11 (orange) and gp12 (purple). (Courtesy of Thomas Goddard, University of San Francisco.)

induce the tail sheath to contract. In this process, the tail tube is driven through the cell envelope to allow the DNA to be injected. The baseplate contains proteins with special functions for this process. One of these proteins has a domain with an enzymatic activity that degrades the bacterial cell wall. This domain (Figure 18.12) is similar to other lysozymes found in most organisms. The phage genome also codes for a lysozyme (T4 lysozyme), but this enzyme is used at the lysis of the bacterium to allow the phage particles to leave.

The structures of several proteins in the baseplate are known. The detailed X-ray structures of the proteins have then been fitted to cryo-EM maps of the baseplate (Figure 18.13).

Recommended Reading

Original Articles

Abad-Zapatero C, Abdel-Meguid SS, Johnson JE et al. (1980) Structure of southern bean mosaic virus at 2.8 Å resolution. Nature 286: 33-39.

Bullough PA, Hughson FM, Skehel JJ, Wiley DC. (1994) Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* **371**: 37–43.

Caspar DLD, Klug A. (1962) Physical principles in the construction of regular viruses. Cold Spring Harb Symp Quant Biol. 27: 1-24.

Hogle JM, Chow M, Filman DJ. (1985) Three-dimensional structure of poliovirus at 2.9 Å resolution. Science 229: 1358-1365.

- Valegård K, Murray JB, Stockley PG, et al. (1994) Crystal structure of an RNA bacteriophage coat protein-operator complex. Nature 371: 623-626.
- Wilson IA, Skehel JJ, Wiley DC. (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* **289**: 366–373.
- Zhang X, Ge P, Yu X, et al. (2013) Cryo-EM structure of the mature dengue virus at 3.5 Å resolution. Nat Struct Mol Biol 20: 105-110.

Review Articles

Harrison SC. (2015) Viral membrane fusion. Virology 479–480: 498–507.

Hogle JM. (2002) Poliovirus cell entry: Common structural themes in viral cell entry pathways. Ann Rev Microbiol 56: 677-702.

Johnson JE, Speir JA. (1997) Quasi-equivalent viruses: A paradigm for protein assemblies. J Mol Biol 269: 665-675.

Smith AE, Helenius A. (2004) How viruses enter animal cells. Science 304: 237–242.

Bioinformatics Tools in Structural Biology

Many methods have been developed as tools for the prediction of folds and functions of macromolecules. In this chapter, we describe the basis of some of these methods, their use and their limitations.

19.1 Structural Comparisons and Classification of Folds

19.1.1 Methods for Structural Comparisons

Structural alignment methods can be used to compare structures and find similarities or differences that are difficult to detect by eye. Similarities may help to reveal unknown evolutionary relations and functional similarities. Sequence alignments based on structural superposition can give useful information about relationships, especially when the level of sequence similarity is low.

Alignment of structures is not as straightforward as sequence alignment. Pairwise alignment can be based on a preconception of which residues should be superimposed in the two structures. If the proteins have obvious sequence similarities, the initial correspondence of the residues can be obtained by sequence alignment. Using the $C\alpha$ coordinates of these residues, a transformation can be calculated that optimally superimposes one of the molecules onto the other as a rigid body. The residues included in the superimposition can be modified using the preliminary alignment to include new residues that are within a suitable threshold distance from each other. After a number of iterations, the result is a superposition that includes a fraction of the residues in the two structures, and

When there is no obvious initial superposition, or when a structure needs to be compared to a whole database of protein structures, a more general method is needed. Since there may be gaps and insertions in the final alignment, a simplified description of the protein folds is required that allows a comparison of elements of the structure rather than the complete molecules. One such description is the distance matrix, which contains the distance between each pair of residues. The result can be plotted as in Figure 19.1, where the proximity of the $C\alpha$ atoms is shown. Helices are seen as thick sections along the diagonal. β -strands are identified as lines perpendicular to the diagonal (antiparallel strands) or parallel to the diagonal (parallel strands). Proximity of secondary structure elements is manifested as contacts at somewhat longer distances. It is easy to see that similar structures will have similar distance matrices. Some methods to find superpositions use pairs of segments of the chains and compare their distance plots. In the most popular method, Dali, these segments are pairs of hexapeptides from both proteins. A similarity score is calculated for the two pairs, and suitable hits are combined to extend the short fragments into a longer alignment.

An alternative simplified representation, used by many fold comparison programs, is to consider only helices and strands and to look for subsets of such secondary structure elements that have similar orientations and directionality.

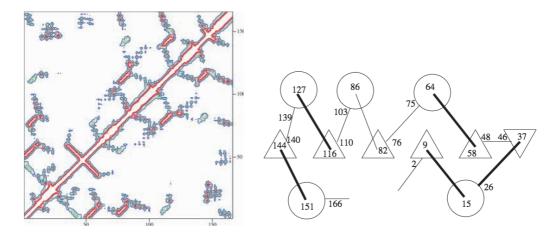


Fig. 19.1 • *Left*: A distance plot for the protein Ras (Section 8.3.3). Contoured regions indicate $C\alpha$ carbon atoms close in space. The plot suggests that there are five helices in the protein (thick segments along the diagonal), five parallel strands and one antiparallel. The Ras protein is partly built up of $\beta\alpha\beta$ units. *Right*: A folding diagram of Ras with an indication of the amino acid residue numbers beginning and ending each secondary structure element (helices: circles; β-strands: triangles). The features of the distance plot can be related to the folding diagram.

19.1.2 Databases of Folds

19.1.2.1 The protein data bank

The major source of information about protein conformation is the Protein Data Bank, which is maintained by wwPDB, an international consortium consisting of RCSB PDB (USA), PDBe (UK) and PDBj (Japan). This database stores experimentally-determined atomic coordinates of proteins and nucleic acid molecules. The information is mainly obtained from structure determination by X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy. The database contains essentially all the detailed structural information on macromolecules that has been generated with these techniques. The actual database consists of the information deposited by the experimentalist as well as annotation by the staff of wwPDB. In addition, the experimental diffraction data are frequently deposited in the case of X-ray structures. An example of a PDB entry is found in Figure 19.2.

19.1.2.2 Quality of experimental models

PDB entries (Figure 19.2) are models of the molecules based on interpretation of experimental data. The coordinates have a statistical error, but unfortunately there are no error estimates associated with the interpretation. The errors are of two kinds: random errors resulting from the limitations of the experimental data, and local or global errors due to various kinds of mistakes by the experimentalists. Before their release, the deposited coordinates are subjected to a number of tests to check for errors or unusual features. This process is called *validation*, and it gives important hints about the correctness or quality of the model as a whole and of individual residues. The most important results of the validation procedure are either included in the header of the entry or available through links. Each one of the entries has references to relevant publications, the sequence of the protein, and a description of the methodology used to determine the conformation, which allows the expert user to judge the quality of the information.

A very useful analysis is the Ramachandran plot (see Section 2.2.4), which is a scatter plot of the conformational torsion angles of the main chain of all residues in a protein (Figure 19.4). A good model will have almost all of these angles in the preferred regions of the plot. If this is not the case, one has to suspect that the model is at least partly incorrect.

For structures determined by X-ray crystallography, the *resolution* of the data is a good indicator of the quality of the resulting model, both with respect to random errors and the risk of mistakes. The resolution measures the level of detail of the diffraction data that was available from the experiment. A low number means high resolution, where data useful for defining the fine details of the structure were available. If the resolution is about 2 Å or higher, the models can normally be expected to be reliable. In some cases, the resolution is as high as 1 Å or even better, and such "atomic resolution" models are

```
HEADER
              VIRUS/VIRAL PROTEIN
                                                                           05-APR-05 1ZA7
              THE CRYSTAL STRUCTURE OF SALT STABLE COWPEA CHOLOROTIC
             2 MOTTLE VIRUS AT 2.7 ANGSTROMS RESOLUTION.
COMPND MOL ID: 1;
COMPND
             2 MOLECULE: COAT PROTEIN;
COMPND 3 CHAIN: A, B, C;
COMPND
             4 SYNONYM: CAPSID PROTEIN, CP;
COMPND
             5 ENGINEERED: YES;
COMPND
             6 MUTATION: YES
SOURCE
             MOL ID: 1;
           2 ORGANISM SCIENTIFIC: COWPEA CHLOROTIC MOTTLE VIRUS;
SOURCE
SOURCE 3 ORGANISM COMMON: VIRUS;
SOURCE 4 GENE: RNA4;
SOURCE 5 EXPRESSION SYSTEM: VIGNA UNGUICULATA;
SOURCE 6 EXPRESSION SYSTEM COMMON: COWPEA;
SOURCE 7 EXPRESSION SYSTEM STRAIN: CALIFORNIA BLACKEYE;
SOURCE 8 EXPRESSION SYSTEM VECTOR TYPE: RNA
KEYWDS MUTANT VIRUS CAPSID STRUCTURE, ICOSAHEDRAL PARTICLE,
KEYWDS 2 STABLIZING MUTATION, STABLE MUTANT, BETA HEXAMER, BETA
KEYWDS 3 BARREL, BROMOVIRUS, POINT MUTATION
EXPDTA X-RAY DIFFRACTION
AUTHOR B.BOTHNER, J.A. SPE
REVDAT 1 21-MAR-06 1ZA7
              B.BOTHNER, J.A. SPEIR, C.OU, D.A. WILLITS, M.J. YOUNG, J.E. JOHNSON
                  AUTH J.A.SPEIR, B.BOTHNER, C.QU, D.A.WILLITS, M.J.YOUNG,
JRNL
JRNL
                 AUTH 2 J.E.JOHNSON
                          ENHANCED LOCAL SYMMETRY INTERACTIONS GLOBALLY
JRNL
                 TITL
JRNL
                TITL 2 STABILIZE A MUTANT VIRUS CAPSID THAT MAINTAINS
               TITL 3 INFECTIVITY AND CAPSID DYNAMICS
JRNL
                                                                         V. 80 3582 2006
               REF
                           J. VIROL.
                REFN ASTM JOVIAM US ISSN 0022-538X
JRNL
ATOM 1 N GLN A 40 127.326 141.523 188.649 1.00 78.03 ATOM 2 CA GLN A 40 126.941 142.963 188.796 1.00 78.71 ATOM 3 C GLN A 40 126.007 143.163 190.001 1.00 77.96 ATOM 4 O GLN A 40 125.985 142.326 190.932 1.00 79.06 ATOM 5 CB GLN A 40 126.243 143.450 187.516 1.00 78.74 ATOM 6 CG GLN A 40 124.899 142.758 187.236 1.00 79.59 ATOM 7 CD GLN A 40 124.192 143.322 186.009 1.00 79.24 ATOM 8 OE1 GLN A 40 124.192 143.322 186.009 1.00 79.24 ATOM 8 OE1 GLN A 40 124.588 143.058 184.869 1.00 81.35 ATOM 9 NE2 GLN A 40 123.138 144.104 186.239 1.00 79.15 ATOM 10 N GLY A 41 125.239 144.262 189.970 1.00 76.76 ATOM 11 CA GLY A 41 124.308 144.563 191.051 1.00 72.98 ATOM 12 C GLY A 41 122.914 144.020 190.777 1.00 69.88 ATOM 13 O GLY A 41 121.981 144.798 190.541 1.00 69.88
                                                                                                                   N
                                                                                                                   C
             13 O GLY A 41
                                             121.981 144.798 190.541 1.00 69.83
ATOM
```

Fig. 19.2 ■ The beginning of the header and the first part of the list of atomic coordinates (labeled "ATOM") of a PDB entry (1ZA7). The format here (PDB format) is one of the available formats of an entry. The coordinates of the first two residues in the entry (Gln 40 and Gly 41 from chain A) are given. The first three columns of numbers are the coordinates of the atoms in Å in relation to a suitable coordinate system. The fourth column of numbers is the occupancy (1.0 for full occupation is normal). In high-resolution structures, some side chains may be modeled in alternative conformations, and multiple coordinates for atoms are given their respective occupancy. In addition, bound molecules may have partial occupancy indicated by this number (for example, if a ligand was bound in only 50% of the protein molecules in the crystal, the occupancy for that ligand would be 0.5). The second to last column gives an estimate of the thermal disorder or the amount of movement of those atoms (the B-factors, see Figure 19.3).

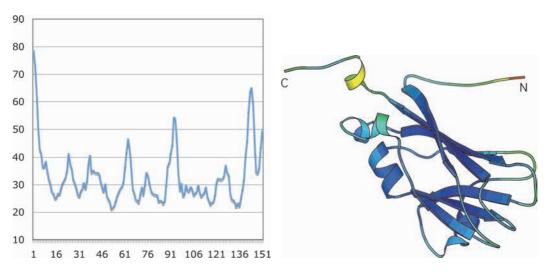


Fig. 19.3 ■ Left: A plot of the B-factors of the $C\alpha$ atoms of the capsid protein of a plant virus (one of the three identical chains in the entry 1ZA7). Right: A schematic drawing of the protein colorcoded according to the B-factor (low: blue, high: red). The B-factors are normally higher in surface loops and at the termini of the chain.

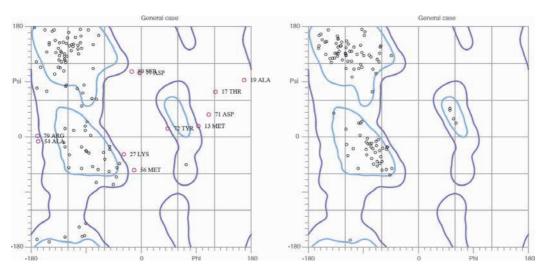


Fig. 19.4 ■ The Ramachandran plot of two different models of the same protein that have been deposited in the PDB. Residues marked by red rings are outside the allowed areas. The plot to the *left* suggests that the coordinates are inaccurate.

very accurate and contain detailed information about hydrogen positions and alternate conformations of side chains. Models at 3-4 Å resolution can still be correct and useful, provided that special care has been taken during the structure determination process. Table 19.1 gives estimates of what can be identified at increasing resolutions.

TABLE 19.1 Estimate of the Resolution Needed to Identify Some Molecular Features of **Proteins and Nucleic Acids**

| Protein | Nucleic Acid | Resolution (Å) |
|--|--|----------------|
| General shape | Distinguish nucleic acid from protein | 20 |
| | Double helix | 12 |
| lpha-helix | Single strand | 9 |
| β-sheet | Stacked base pairs, phosphate groups and riboses | 4 |
| Large side chains | | 3.5 |
| Shaped bulbs of density for small side chains | Purines and pyrimidines distinguished | 3.2 |
| Conformation of side chains | | 2.9 |
| Carbonyl groups, peptide planes | Individual bases | 2.7 |
| | Puckering of sugar | 2.4 |
| Puckering of Pro residues, holes in aromatic rings | | 2.0 |
| Individual atoms | Individual atoms | 1.5 |

The electron density itself can be very informative and is available for many entries through the EDS, the electron density server. With this tool it is easy to see the experimental evidence for side chain positions and ligand conformations. The experimental data that can be used to recreate the electron density is also available from the PDB website.

R-values describe how well the atomic model explains the observed diffraction data. The most important one is the R_{free} , which compares the model with a small fraction of diffraction data that have not been used for the refinement. That is, the $R_{\rm free}$ data set has not been used to create the model or optimize the model to fit best with the diffraction data. A lower R_{free} generally indicates a more accurate model. The R-values improve (decrease) with increasing overall resolution since the modeling becomes more accurate. Therefore, structures modeled to 1.5 Å resolution should have lower R_{free} values than structures modeled to 3.5 Å resolution. With more than 100 000 crystal structures now deposited in the Protein Data Bank, we can compare the R_{free} value of a given structure to other Protein Data Bank entries of a similar resolution. This will give us an indication of the quality of the structure in question, as a significantly higher than normal R_{free} for that resolution will suggest lower quality of the model or the experimental data used to generate the model. This comparison is done automatically by the Protein Data Bank and is displayed in the online entry for every X-ray crystal structure.

Another phenomenon that can significantly improve the electron density (and therefore the model that is built from it) is the possibility of averaging identical molecules. Virus molecules with large numbers of identical protein subunits in different orientations

give excellent options for averaging and improvement of the electron density. This allows modeling of proteins at relatively low resolution.

The various wwPDB nodes (RCSB, MSD, PDBj) each offer their own set of search facilities. The coordinates are available for downloading for further analysis. In addition, the various coordinate sets (entries) have links to several other databases: sequence databases, graphical representations of the structures and others.

In February 2016, there were about 115 000 entries in the PDB, but many of them are mutants of the same protein or complexes of a protein with various ligands. For example, there are more than 400 different entries for mutants of lysozyme from bacteriophage T4 (a favorite for the study of protein folding and stability) and more than 700 entries for various inhibitor complexes of the HIV protease.

19.1.2.3 Hierarchical fold databases: SCOP and CATH

A number of databases are based on the information collected in the Protein Data Bank. These databases contain images of molecules, sequences and other types of information. Two databases organize the available structural information in a hierarchical arrangement. These databases are useful for the analysis of evolutionary relationships as well as for functional comparisons.

In the SCOP (Structural Classification Of Proteins) database developed by A. Murzin and his coworkers (UK), all known protein structures are sorted according to their fold. This database is mainly based on manual classification of the protein folds and therefore the classifications are subjective. On the other hand, the accumulated knowledge of protein conformation and the details of the evolution of specific proteins that is used for the classification are not easily incorporated in a computer program for automatic fold classification.

The levels of the hierarchy (Table 19.2) are classes, folds, superfamilies, families and domains (the individual proteins or protein domains). The main classes are all alpha, all beta, alpha/beta and alpha+beta. The various domains within a family are assumed to be homologous, i.e. to have a common ancestor from which they have diverged. The hypothesis regarding homology of protein domains with similar structures is based on sequence and/or functional similarity. Proteins within one superfamily have the same fold and

| TABLE 19.2 | The Hierarchy | Levels in SCOP |
|-------------------|---------------|----------------|
|-------------------|---------------|----------------|

| | <u> </u> |
|-------------|--|
| Class | Main grouping (alpha, beta, etc.) |
| Fold | Type of fold |
| Superfamily | Structural similarity and common ancestry likely |
| Family | Common ancestry |
| Domain | Protein |
| | |

a related function and therefore they also probably have a common ancestor, but they differ too greatly in sequence or function to allow a conclusive decision about homology. At the next level, fold, the proteins have the same topology (Section 3.2.3), but there is no evidence of an evolutionary relationship between superfamilies except the limited structural similarity.

The SCOP database classifies domains rather than entire proteins. This means that multi-domain proteins are divided into their constituent domains. This is very useful in the many cases where one kind of domain is shared by different proteins.

A similar structure database is CATH, developed by C. Orengo and J. Thornton (UK). It orders every known protein structure hierarchically in classes, architectures, topologies (fold families), superfamilies and sequence families (CATH stands for Class, Architecture, Topology, Homologous superfamily). The topologies of the CATH database correspond essentially to the folds of the SCOP classification, but it introduces a level between Class and Fold (Table 19.2), where proteins with the same spatial arrangement of secondary structure elements are grouped. Architectures defined in CATH are shown in Figures 3.2-3.4.

The classification of structures in CATH is to a large extent done by automated procedures. The automatic procedure starts with the definition of domains in a protein. A number of procedures for doing this automatically have been developed. Due to the differences in classification procedures, some proteins are grouped differently in SCOP and CATH.

19.1.2.4 Other structural classifications

There are structure databases that are not hierarchical but are based on automated clustering of known structures. One of these is the Dali database, which uses the program Dali to compare and classify protein folds (Table 19.3).

The procedure is completely automatic, in contrast to the procedures used in CATH and SCOP. For all structures in a representative set with less than 90% sequence identity, the Dali database contains a list of all structurally similar molecules in the PDB. This list contains information about the degree of structural and sequence similarity between a given protein and the other entries. Proteins in the list are sorted according to their structural similarity by a Z-score. The Z-score is calculated as the number of standard deviations above the average structural similarity of random proteins of the same size. Proteins with a Z-score of less than 2.0 are regarded as dissimilar and not included in the list, but there is no precise Z-score limit that can be used as a threshold for an evolutionary relationship. Since the list is based purely on a computed similarity score, any functional similarity can be an important additional factor for a conclusion about evolutionary relationships.

There are other programs for structure comparisons. NCBI uses the program VAST (Vector Alignment Search Tools) for comparisons of structures and its structure database MMDB contains structure neighbors to all PDB entries. PDBe has a server for structure comparisons called PDBeFold.

TABLE 19.3 Structural Neighbors of Hexokinase (1QHA)

| Chain ^a | Z-Score ^b | %ID ^c | LALI ^d | RMSD ^e | Description |
|---------------------------|----------------------|------------------|--------------------------|-------------------|---|
| 1qhaA | 69.1 | 100 | 903 | 0.0 | Hexokinase (human) |
| 1bg3B | 62.4 | 92 | 883 | 0.9 | Hexokinase (rat) |
| 1bdg | 55.4 | 45 | 439 | 1.5 | Hexokinase (blood fluke) |
| 1v4sA | 55.2 | 54 | 444 | 1.5 | Glucokinase isoform 2 |
| 1hkg | 40.7 | 15 | 432 | 2.3 | Hexokinase (yeast) |
| 1ig8A | 39.2 | 34 | 438 | 3.4 | Hexokinase PII (yeast) |
| 1xc3A | 19.8 | 10 | 265 | 3.1 | Putative fructokinase |
| 1q18A | 18.5 | 13 | 280 | 4.5 | Glucokinase |
| 1woqA | 17.7 | 16 | 239 | 3.0 | Inorganic polyphosphate/ATP glucomannokinase |
| 1e4gT | 14.8 | 10 | 261 | 4.0 | Cell division protein FtsA (<i>Thermatoga maritima</i>) |
| 1yagA | 14.8 | 9 | 266 | 4.1 | Actin (yeast) |
| 2btfA | 14.7 | 9 | 269 | 4.1 | Actin (cow) |
| 1jcfA | 14.7 | 11 | 258 | 3.6 | Rod shape-determining protein MreB (Thermatoga maritima) |
| 1dkgD | 14.6 | 10 | 267 | 3.9 | Nucleotide exchange factor GrpE (E. coli) |
| 1s3xA | 14.1 | 10 | 269 | 4.3 | HSP70 (human) |
| 1mwkA | 13.5 | 10 | 250 | 3.5 | Plasmid segregation protein ParM (E. coli) |
| 1huxA | 13.3 | 14 | 224 | 3.8 | Activator of hydroxyglutaryl-CoA dehydratase (<i>A. fermentans</i>) |
| 1nbwA | 12.2 | 10 | 265 | 4.0 | Glycerol dehydratase reactivase α subunit |
| 1tuuA | 11.7 | 10 | 252 | 4.2 | Acetate kinase |
| 1glfY | 10.6 | 14 | 245 | 3.7 | Glycerol kinase |
| 1hjrA | 6.2 | 5 | 123 | 4.0 | Holliday junction resolvase RuvC |
| 1c0mC | 4.3 | 5 | 87 | 2.9 | Integrase (Rous sarcoma virus) |

^a PDB code followed by the name of the chain (A, B, ...). The table shows the structural neighbors of chain A in entry 1QHA (hexokinase).

^bThe Z-score indicates the similarity of the structure to the input structure. The list is abbreviated. The top of the list includes several hexokinase structures. Relatively high scores are also indicated for actins (see Section 15.1) and actin-related proteins (MreB, FtsA, ParM) and heat-shock proteins (HSP70, see Section 12.1.2.3). Both these groups of proteins are partly similar to hexokinase, although the sequences are very different. A common function (ATP binding and hydrolysis) suggests an evolutionary relationship. The last two chains in the list are included as examples of proteins with only local structure similarity.

^cPercentage identities among the aligned residues.

^dNumber of residues aligned.

 $^{^{\}rm e}$ R.M.S. distance (in Å) of the C α atoms of the chains after superposition.

19.2 **Prediction of Protein Conformation**

19.2.1 Secondary Structure Prediction

19.2.1.1 Basis of secondary structure prediction methods

The realization that protein sequences contain all the information necessary to dictate the folding of a protein has inspired scientists to try to predict protein conformation directly from the sequence. Despite decades of effort, this goal has not yet been achieved except for some small proteins with simple folds. The lack of success is mainly due to two phenomena. First, even a short protein has an enormous number of possible conformations (Section 3.1.1.1). Even with modern computers, it is far from possible to generate all these conformations and calculate their energy, even if the task is simplified by allowing only a small number of arrangements of every amino acid residue. Second, the free energy of the system (the protein and the surrounding solvent) has to be calculated using simplified potential functions, which may not account correctly for all forces in the system and therefore may not be able to identify the correct minimum. In addition, the folded conformation is only marginally more stable than the unfolded state (Section 3.1.1.1).

At an early stage, it was realized that the problem could be simplified by dividing the prediction into two stages: first, the secondary structure elements of the protein are predicted, and the strands and helices are assembled into a fold. In a way, this approach mimics one possible mechanism of protein folding in which local secondary structure is formed first and these subsequently arrange themselves to achieve the folded conformation.

The secondary structure of a segment of the polypeptide chain in a folded protein depends on the amino acid sequence. One approach for prediction is to use the tendency of amino acids to prefer one kind of secondary structure conformation. Secondary structure prediction methods have been based on such propensities for amino acid residues to form β -strands, helices or turns. These propensities have been derived from studies of the conformation of small peptides in solution, or from statistical analysis of the occurrence of certain residues in the various types of secondary structure in proteins of known structure. The properties of some amino acid residues (most notably Gly and Pro residues) in relation to the fold is discussed in Section 2.2.4.

19.2.2 Prediction Methods

19.2.2.1 Chou-Fasman method

The first widely used procedure for secondary structure prediction was the Chou-Fasman method. This method was based on α -helical and β -strand propensities of all 20 amino acids, which are classified as helix-forming or -breaking and strand-forming or -breaking (Table 19.4). These propensities are based on a statistical analysis of the occurrence of various residues in secondary structure elements of known protein structures. The method

TABLE 19.4 Chou–Fasman Propensities for α-Helix, β-Strand and Turn

| Helix | | | | Strand | | | Turn | |
|-------|------|---|-----|--------|---|-----|------|--|
| Glu | 1.51 | Н | Val | 1.70 | Н | Asn | 1.56 | |
| Met | 1.45 | Н | Ile | 1.60 | Н | Gly | 1.56 | |
| Ala | 1.42 | Н | Tyr | 1.47 | Н | Pro | 1.52 | |
| Leu | 1.21 | Н | Phe | 1.38 | h | Asp | 1.46 | |
| Lys | 1.16 | h | Trp | 1.37 | h | Ser | 1.43 | |
| Phe | 1.13 | h | Leu | 1.30 | h | Cys | 1.19 | |
| Gln | 1.11 | h | Cys | 1.19 | h | Tyr | 1.14 | |
| Trp | 1.08 | h | Thr | 1.19 | h | Lys | 1.01 | |
| Ile | 1.08 | h | Gln | 1.10 | h | Gln | 0.98 | |
| Val | 1.06 | h | Met | 1.05 | h | Thr | 0.96 | |
| Asp | 1.01 | I | Arg | 0.93 | i | Trp | 0.96 | |
| His | 1.00 | I | Asn | 0.89 | i | Arg | 0.95 | |
| Arg | 0.98 | i | His | 0.87 | i | His | 0.95 | |
| Thr | 0.83 | i | Ala | 0.83 | i | Glu | 0.74 | |
| Ser | 0.77 | i | Ser | 0.75 | b | Ala | 0.66 | |
| Cys | 0.70 | i | Gly | 0.75 | b | Met | 0.60 | |
| Tyr | 0.69 | b | Lys | 0.74 | b | Phe | 0.60 | |
| Asn | 0.67 | b | Pro | 0.55 | В | Leu | 0.59 | |
| Pro | 0.57 | В | Asp | 0.54 | В | Val | 0.50 | |
| Gly | 0.57 | В | Glu | 0.37 | В | Asn | 1.56 | |

H = strong former, h = former, I, i = indifferent, B = strong breaker, b = weak breaker. (From Chou PY, Fasman GD (1978) Ann Rev Biochem 47: 251-276.) A helix or a strand is predicted if the character of the residues is favorable in a window of six or five residues, respectively. The helix or strand is extended until the average propensity in a window of four residues falls below 1.

consists of a number of rules for helix and strand formation and extension using average propensities in a short segment. The methods predict the secondary structure as one of three states: helix, strand or loop, where the loop state includes all conformations that are not helices or β-strands. The simple Chou–Fasman prediction method can be performed with a pencil and a piece of paper, although a programmed version is normally used.

19.2.2.2 Neural network methods

A number of methods that show significantly better results than the early methods have been developed. The improvement is due in part to the use of multiple aligned sequences for the prediction. The pattern of substitutions and insertions/deletions in homologous

proteins contains information about features that are of importance for the fold of a protein. Insertions and deletions are, for example, rare in secondary structure elements and conserved glycine and proline residues are likely to be found in a turn.

One example of such a program is *PHDsec*, which predicts secondary structure based on a two-layered feed-forward neural network.

A neural network is a computer program that can incorporate knowledge about relations between various types of input and output information. The information is stored as parameters (weights) in the program. In the case of a neural network for secondary structure prediction, the aim is to connect sequence patterns with the secondary structure to which they have been observed to correspond. The network consists of a number of layers (input layer, hidden layer(s), output layer), and the information goes from one layer to the next (feed-forward). Aligned homologous sequences of known structures are used to "train" the network. In this procedure, the relation between sequence and observed secondary structure is used to calculate parameters. These parameters implicitly describe the probability that a certain residue will result in a certain secondary structure and can then be used to predict the secondary structure of the aligned sequences of the unknown protein.

The neural network in PHD (Figure 19.5) has one layer which links sequence patterns to secondary structure (sequence-structure layer). In the training step, the occurrence of various residues in a window of 13 amino acids is correlated with the secondary structure of the central residue. In the second step (structure-structure layer), the output from the first layer in a window of 17 residues is used to predict the secondary structure of the central residue. In this case, the network will be trained not to predict unrealistically short segments of secondary structure.

When the neural network has been trained using several proteins of known structure, it can be used for prediction. One of the main features of PHDsec is that the search sequence is first used to find homologous sequences (using the sequence alignment tool BLAST) that are aligned to the search sequence. The aligned sequences are fed into the network and used for prediction.

19.2.2.3 Accuracy assessment

The standard method for assessing the accuracy of secondary structure prediction methods is to predict the secondary structure for a number of proteins of known structure and to calculate the Q number, the percentage of residues assigned to the correct class. For a completely random prediction in three classes (Q3 score, classes helix, sheet and other) this number will be 33% if the proteins have equal amounts of the three categories of conformation. The best programs, based on neural networks and other methods can reach close to 80% accuracy when multiple sequences are available. It would be very useful if the quality of a specific prediction could be estimated. Unfortunately, there is no good method for doing this.

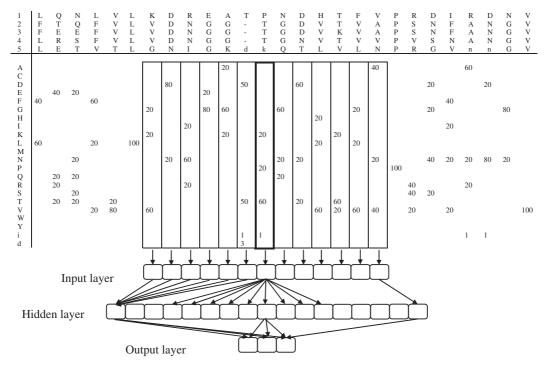


Fig. 19.5 ■ Schematic drawing of part of the neural network in PHDsec. The protein is aligned to four more sequences, and the sequence information in a window of 13 residues is put into the input layer of $13 \times 24 = 312$ units (one for each amino acid and some coding for ends and gaps). A hidden layer of 17 nodes connects the input layer to the output layer, with one node each for the probability of helix, strand and loop. A second network takes the output from the first network to improve the local prediction obtained in this step.

In some cases, many of the secondary structure elements are predicted correctly, but their start and end are not. In fact, there is often an ambiguity in the definition of the beginning and end of secondary structure elements even when the structure is known.

The best secondary structure prediction methods are now relatively accurate. Some procedures combine a number of different methods to take advantage of differences in the ability of these methods to find a good prediction. Secondary structure prediction methods, however, are always based on the local sequence, but the true folding and thus the secondary structure depends on non-local interactions. There are cases where identical sequences of upto eight residues form completely different structures, which shows that non-local effects can be crucial for the conformation of a peptide (chameleon sequences; Figure 19.6).

19.2.2.4 *CASP*

The best test of prediction methods is to try them on proteins for which the structure is unknown. Such tests are the focus of CASP (Critical Assessment of Techniques for Protein

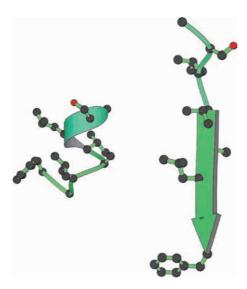


Fig. 19.6 ■ The same octapeptide (GSLVALGF) can have both α -helical (*left*) and β -strand (*right*) conformation in phosphatidylinositol 3-kinase (*left*, PDB: 1PHT) and a chymotrypsin inhibitor (*right*, PDB: 1WBC).

Structure Prediction). Scientists involved in modeling and method development are invited to model protein structures that are about to be solved by experimental methods. The models are submitted and compared to the experimentally determined structures when these become available and the results are discussed at a meeting every second year. CASP has been important for the development of prediction methods in all fields discussed in this chapter.

19.2.3 Prediction of Other Local Properties

19.2.3.1 Prediction of topology of transmembrane proteins

Throughout Section 4.6 in Chapter 4, we saw how different regions of helical membrane proteins have different sequence characteristics (summarized in Figure 19.7). Given these amino acid biases, it follows that we should be able to exploit this knowledge to predict a protein's topology from its sequence. Attempts towards this end have been underway for more than 20 years, and initially relied upon the two hallmark features of helical membrane proteins: extended stretches of hydrophobic amino acids (Section 4.6.2) and the positive-inside rule (Section 4.6.6.1). Many different topology prediction programs have now been developed, some of which are listed at the end of this chapter. Most modern programs are based upon "machine learning" methods, in which databases of membrane proteins of known structure or topology are used to identify statistically-significant amino acid biases in different areas (such as the hydrophobic core, interfacial regions, and

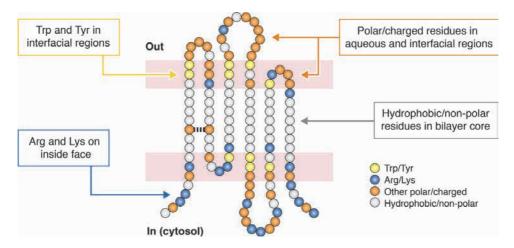


Fig. 19.7 Amino acid location biases in helical membrane proteins. The interfacial regions are represented by red bars. The dashed line represents a hydrogen bond between membrane-embedded polar residues.

the inner and outer loops). These location biases can then be used to analyze the sequence of a protein of unknown structure and generate the most probable topological model. Figure 19.8 shows the predicted topology of the protein Rhodopsin, compared with its experimentally determined structure.

In addition to standard amino acid analyses, some topology prediction programs can incorporate other information to increase their accuracy. For example, predictions can be made from multiple sequence alignments rather than one single sequence, or prior topological information — perhaps from experimental data — can be used to constrain regions of a sequence to a particular location. These and other methodological advances have greatly improved the success of membrane protein topology predictions, which are generally of good accuracy. However, there still remain some challenges. For example, predictions can be misled by the presence of reentrant loops (Section 4.6.4.3), or if a cleavable N-terminal membrane-targeting signal sequence (Section 4.4.2) is mistaken for a transmembrane helix (since both are hydrophobic in nature). To overcome this common pitfall, some topology prediction programs now separately detect and account for such signal sequences.

19.2.3.2 Prediction of disordered segments

Most proteins have an ordered conformation, but various kinds of flexibility are often important for their function. Many proteins contain disordered segments. In some, the disordered parts do not have any obvious functions, but in other cases the disordered segments have important roles. They may become ordered only when interacting with

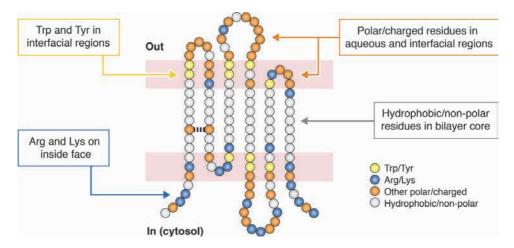


Fig. 19.7 Amino acid location biases in helical membrane proteins. The interfacial regions are represented by red bars. The dashed line represents a hydrogen bond between membrane-embedded polar residues.

the inner and outer loops). These location biases can then be used to analyze the sequence of a protein of unknown structure and generate the most probable topological model. Figure 19.8 shows the predicted topology of the protein Rhodopsin, compared with its experimentally determined structure.

In addition to standard amino acid analyses, some topology prediction programs can incorporate other information to increase their accuracy. For example, predictions can be made from multiple sequence alignments rather than one single sequence, or prior topological information — perhaps from experimental data — can be used to constrain regions of a sequence to a particular location. These and other methodological advances have greatly improved the success of membrane protein topology predictions, which are generally of good accuracy. However, there still remain some challenges. For example, predictions can be misled by the presence of reentrant loops (Section 4.6.4.3), or if a cleavable N-terminal membrane-targeting signal sequence (Section 4.4.2) is mistaken for a transmembrane helix (since both are hydrophobic in nature). To overcome this common pitfall, some topology prediction programs now separately detect and account for such signal sequences.

19.2.3.2 Prediction of disordered segments

Most proteins have an ordered conformation, but various kinds of flexibility are often important for their function. Many proteins contain disordered segments. In some, the disordered parts do not have any obvious functions, but in other cases the disordered segments have important roles. They may become ordered only when interacting with

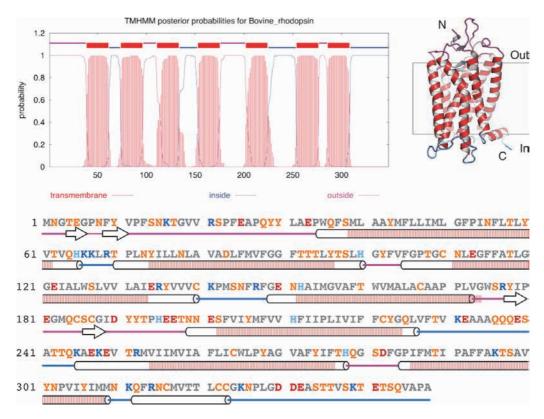


Fig. 19.8 ■ The predicted topology of Rhodopsin. *Top left*: The graphical output from the program TMHMM (www.cbs.dtu.dk/services/TMHMM), predicting the topology of bovine Rhodopsin from its sequence. Bottom: The sequence of rhodopsin (UniProt ID P02699). The locations of the secondary structure elements from the experimentally-determined 3D structure (top right, PDB: 1FF8) are shown below the sequence as cylinders (helices) and arrows (sheets). Red shading corresponds to the transmembrane regions predicted by TMHMM, demonstrating close agreement between the predicted structure and the true structure. Inside and outside loops are colored blue and purple, respectively. The prevalence of arginine and lysine residues on the inside loops is also evident from the sequence, consistent with the positive-inside rule (Section 4.6.6.1).

other molecules, which is a property that can be used for achieving optimal binding affinity and specificity. This is further discussed in Section 3.2.4.

Prediction of disorder is of interest, both for functional analyzes and for the choice of suitable fragments for structural studies. Disordered regions in proteins are sometimes found as "low-complexity" regions, where the distribution of amino acids is not the same as in globular proteins. These regions may have unusually high frequency of residues like lysine, glutamine, glutamic acid, glycine or proline, and a lack of large non-polar residues.

Prediction of disordered segments can be achieved with neural networks that are trained with segments or whole proteins that are disordered based on experimental evidence from X-ray crystallography and other methods. The procedure is similar to what

has been described for prediction of secondary structure (Section 19.2.2.2). Examples of such methods are DISOPREP and SPINE-D. There are also databases of disordered proteins or protein segments, for example, DisProt and IDEAL.

19.3 **Modeling of Protein Tertiary Structure**

19.3.1 Comparative Modeling

The large number of experimentally determined structures makes it possible to model a protein using a suitable known structure as a starting point (homology modeling, comparative modeling or template-based modeling). This method is based on the observation that homologous proteins have similar structures. Such modeling may also be useful when the sequence similarity is low. The first step in the modeling is to find one or several suitable templates of known structure.

19.3.1.1 Searching for homology models using folds — fold recognition

In many cases, the modeling can be based on a protein of known structure with obvious sequence similarities to our protein of interest. For some proteins, however, it is difficult to find a clearly homologous protein of known structure. The most sensitive methods to establish homology based on sequences alone use a sequence profile of a family of proteins. This is the method used in PSI-BLAST. When these methods fail to find a homologous protein, other procedures that use structural information to aid the comparisons can be used. These procedures select the most probable fold for a protein from the database of all known folds. This is called *fold recognition*. These methods have also been labeled *inverse* folding, since you look for the compatibility of a sequence with a specific fold. Fold recognition methods have to find the optimal alignment to the templates, and to calculate a score that is a good estimate of how well the target sequence fits the template. This score is used to decide whether a protein of known fold is useful as a template for modeling or not.

Early methods, called threading methods, used contact potentials to calculate a score for each alignment of the target sequence to each template. These contact potentials took into account the character of the hydrohobic core and the surface. Computing the score for a very large number of alignments takes a long time, and this limitation has led to the development of methods that more efficiently find an optimal alignment.

Profile methods use the powerful dynamic programming method for alignment of a linear description of the template and target. These methods use the amino acid sequences of the target and template and a scoring method that may be general, like the common Many modern methods for fold recognition are relatively difficult to describe in simple terms, but they are able to pick up homology between the sequence of an unknown protein and a distantly related protein of known structure. They are often available as servers that can be used freely. Examples of such servers are Phyre2, I-TASSER, SWISS-MODEL, HH-PRED and Raptor.

These servers also produce models based on the best available known structures. The crucial step in the modeling is the alignment of the target sequence to the template(s). If the sequences are not very similar, the problem is to correctly place deletions and insertions. The alignment can be based on an analysis of many related sequences. Automated procedures may use several alignments and produce models from each of them.

19.3.1.2 Quality of the models

Homology modeling can result in fairly accurate models, especially in cases where the template has a high degree of sequence similarity to the target protein and there are no errors caused by incorrect alignment of the target sequence to the template.

When the sequence similarity is low (below 30%), models based on sequence homology will most likely be partly incorrect. Even if the fold was identified correctly, the difficulties in aligning the sequences correctly make it likely that the sequence will be fitted incorrectly not only in surface loops, but possibly also in secondary structure elements. In addition, when the differences in conformation between the starting model and the true structure are large, the modeling is unlikely to find the correct conformation. Since active sites and other parts of the protein that are important for the function are often more conserved than other regions, such homology models can still be useful for predicting functional properties that can be tested.

The success rate of fold recognition methods has been analyzed in the CASP project. This "competition" may have contributed to the development of the methods, and for a significant fraction of sequences the correct fold can be found also in cases where sequence similarity is hard to detect. These programs are therefore of use because of the possibility of assigning a fold and therefore a tentative function to an unknown protein.

19.3.1.3 Modeling without homology: ab initio modeling

The modeling of tertiary structure without reference to a specific template with similar structure has a long history. These methods have been called *ab initio* methods, new fold prediction methods or free modeling methods. Early methods used simplified descriptions

of the amino acid residues and tried to calculate the conformation of lowest energy using a minimization method. Such procedures attempted to model the folding pathway of a protein and were therefore labeled ab initio folding. These methods have had limited success compared to fold recognition methods, since the original approach suffered from limitations in the potential functions used and from the size of the computational problem. However, folding proteins computationally is still a very active area of research. Molecular dynamics methods can be used to simulate trajectories of protein molecules in solution, but these calculations can only be performed for fractions of the time scales of the complete folding of most proteins.

19.4 **Assignment of Function to Proteins**

19.4.1 Assignment of Function through Sequence Similarity

19.4.1.1 Sequence similarity and homology

The flood of genome sequence data has made it necessary to develop methods to find possible functions of proteins coded for by unknown genes. The main method for assignment of a function to a protein is to establish homology to a protein of known function using sequence alignment methods. The alignment score is a measure of the similarity between two sequences, and with proper statistical treatment it will give an estimate of how likely it is that the alignment score could be obtained by chance. This assumes that truly unrelated sequences have no systematic similarities. Although there are some examples of convergent adaptive changes of a small number of residues in a protein, sequence convergence for whole proteins and domains has not been observed. Significant sequence similarity between proteins or domains can therefore always be used to prove homology.

Although the best way to decide whether two proteins are homologous or not is to do a proper statistical analysis of alignment scores, it is often convenient to use the degree of sequence identity. Comparisons of many sequences and structures show that a 35% sequence identity between two sequences (calculated as the number of identities divided by the total number of amino acids in the shorter of the two sequences) can be regarded as reasonable evidence of homology. For longer sequences, a lower cutoff can be used (Figure 19.9). Below the cutoff lies is a gray zone (sometimes labeled the "twilight zone") where many true homologues are found but it is not possible to decide whether the similarity is due to homology or not. If the level of sequence identity is low, more sophisticated methods, for example, protein-specific profiles or hidden Markov models, can be used to detect significant sequence similarity.

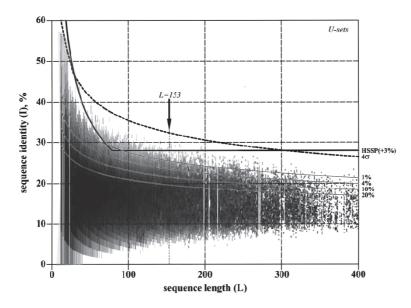


Fig. 19.9 ■ A plot of the degree of protein sequence identity as a function of sequence length for 1.3 million unrelated sequences. The sequences were aligned with the Needleman-Wunsch algorithm with no end gap penalty. Various thresholds are indicated. (Reprinted with permission from Abagyan RA, Batalov S. (1997) J Mol Biol 273: 355-368.)

As discussed in Chapter 3, there are proteins with the same fold but with no other evidence of a common ancestry. This similarity may be due to structural convergence: the similarity may be the result of independent evolution of the same fold from different ancestors.

19.4.1.2 Conserved sequence patterns

In many cases, conserved residues in homologous proteins show a specific pattern. Since functionally important residues tend to be conserved, such patterns can be used to tentatively identify unknown proteins and connect them to a function in the absence of a global sequence similarity. For example, insulin contains a characteristic pattern C-C-{P}-{P}-x-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C, where residues in curly brackets indicate disallowed amino acids and square brackets group together allowed residues. This pattern includes four conserved cysteine residues that form disulfides, and searching for this pattern retrieves about 200 insulin chains (true positives) in SwissProt, but misses 10 chains (false negatives) and finds three proteins that are not insulins (false positives). The usefulness of such a pattern can be described by its selectivity (fraction of hits that are correct) and sensitivity (fraction of the protein searched for that was found by the pattern).

The number of conserved residues in the globin family is very small, and searches for short patterns yield many false positives. A sequence profile covering the complete

sequence and encoding the probability of finding a certain residue at a certain position can be used instead to identify globins with high precision. The sequence is aligned to the profile and a score is calculated. If the maximum score is above a cutoff value, it is classified as a hit. In ProSite, this type of profile searching has identified 738 globins in SwissProt with only one false positive.

19.4.1.3 Do homologous proteins have the same function?

In most cases, significant sequence similarity between proteins suggests that they have the same or similar functions but there are a number of exceptions. The extreme case is "moonlighting", where the same protein fulfills two, or even several different functions in an organism. There are many examples of such proteins, and the functions can be completely unrelated. A well-known example of this phenomenon is that several crystallins, proteins in the eye lens are closely related to metabolic enzymes and some of them have catalytic activity. In evolution, these proteins (for example, quinone reductase) may have been recruited as crystallins because they have suitable properties and can be produced in large amounts to fill the lens. In other cases, the protein can have different functions inside and outside the cell.

Homologous proteins are either orthologs, where the same gene has changed during the evolution of organisms or paralogs, where gene duplication has created two copies of a protein gene in one organism and these have developed into two distinct proteins through mutations. Orthologs can be expected to have similar function, at least for closely related organisms, but paralogs may have more distinct functions. A high degree of sequence similarity for paralogs does not necessarily indicate that the function is very similar. One example is (hen) lysozyme and (goat) α -lactalbumin, which have a sequence identity as high as 45% but very different functions. In this case, the gene duplication probably occurred in a common ancestor to birds and mammals. Most paralogs, however, have been recruited by the organism to perform a function related to the original function.

19.4.1.4 Databases of aligned sequences and sequence patterns

The database InterPro is maintained by the European Bioinformatics Institute in UK. Its goal is to integrate the information about protein families that are found in several other databases. One of these, the *Pfam* database (UK), contains a large number of aligned sequences of domain families. Note that protein domains are organized in Pfam; a single multi-domain protein can therefore be found in several families. The alignment has been made using the hidden Markov model methodology. The hidden Markov model (HMM) for each family is in this case based on an original alignment of a limited number of sequences using Clustal, combined with manual adjustments (seed alignment). Many other sequences have then been added to the families using the HMM by searching in a non-redundant protein sequence database (SwissProt + TrEMBL). In the Pfam database, the domain organization of proteins is shown graphically, and search sequences can be analyzed in the same way.

A number of databases and servers contain information about conserved sequence features (patterns or profiles) in proteins. The ProSite database maintained by the Swiss Institute of Bioinformatics, contains close to 1500 patterns that can be used to identify function of a protein. The database contains a list of sequence database entries (with links) that contain each pattern. True- and false-positives are indicated as well as known false negatives. In addition, the database contains a description of the pattern and the protein family.

19.4.2 Structure-based Function Prediction

19.4.2.1 Structural genomics

Although the main method for the assignment of a function to a protein is finding a significant sequence similarity, there are many cases where no related protein of known function can be found. In cases where no significant similarity at the sequence level is found, one way to identify the function of a gene is to determine the three-dimensional structure of the protein and, from the fold, try to decide what function the protein might have. Although there is in general no relation between the fold and the function of a protein, there are cases where a fold is connected with only one or a few related activities.

19.4.2.2 Function prediction using templates

If the three-dimensional structure is known, prediction of function can be based on local structural similarity. Three-dimensional templates of active sites of enzyme, DNA- and ligand-binding surfaces and reverse templates defined from the target protein itself can be used to find structural similarity between the target protein and proteins of known function. The templates describing the active sites consist of two to five residues, while the other templates consist of three residues. This approach is used in the ProFunc server (EBI, UK), where the template method is combined with sequence comparisons and global structural comparisons to predict the function of a protein of known structure.

19.4.3 Assigning Function by Genome Comparison

The function of unknown genes can also be found indirectly. One way to find out whether two proteins are involved in the same metabolic pathway is to look for a possible gene fusion of the two proteins in another genome. A gene fusion in one organism is a strong indication that the two proteins are physically associated or perform successive steps in another organism. In bacterial genomes, hundreds of pairs of proteins are found as fusion proteins in one or another organism.

Another way to assign a function to a protein is to analyze the phylogenetic profile. There are a number of such genomic-context methods. Pairs or groups of proteins that are all either present or absent in genomes are likely to be physically or functionally associated and conserved gene order or shared regulatory elements indicate that the proteins are related.

Recommended Reading

Bonneau R, Strauss C, Rohl C, et al. (2002) De novo prediction of three-dimensional structures for major protein families. J Mol Biol 322: 65–78.

Bowie JU, Luthy R, Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.

Holm L, Sander C. (1993) Protein structure comparison by alignment of distance matrices. I Mol Biol 233: 123-138.

Laskowski RA, Watson JD, Thornton JM. (2005) Protein function prediction using local 3D templates. J Mol Biol 351: 614-626.

Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.

Orengo CA, Michie AD, Jones S, et al. (1997) CATH — A hierarchic classification of protein domain structures. Structure 5: 1093-1108.

Rost B, Sander C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19: 55–72.

von Heijne G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225: 487-494.

Links to Databases and Servers

PDB at RCSB: www.rcsb.org/pdb/home/home.do

PDB at PDBe: www.ebi.ac.uk/pdbe

PDBj: www.pdbj.org/

SCOP: scop2.mrc-lmb.cam.ac.uk

CATH: www.cathdb.info

Dali: ekhidna.biocenter.helsinki.fi/dali/start

NCBI structural database MMDB: www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

VAST: www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml

CASP: predictioncenter.org/

ModBase: modbase.compbio.ucsf.edu/modbase-cgi/index.cgi

SignalP (distinguishing signal sequences in membrane proteins): cbs.dtu.dk/services/SignalP

SwissModel: swissmodel.expasy.org/

PredictProtein (Server for several prediction programs): www.predictprotein.org/

PHYRE2, fold recognition and other prediction programs: www.sbg.bio.ic.ac.uk/

Membrane-spanning barrels: http://cubic.bioc.columbia.edu/services/proftmb/

InterPro: www.ebi.ac.uk/interpro/

Pfam: pfam.xfam.org

Phobius (topology prediction membrane proteins): phobius.binf.ku.dk

ProSite: prosite.expasy.org

ProFunc: www.ebi.ac.uk/thornton-srv/databases/profunc/

TMHMM (topology prediction membrane proteins): cbs.dtu.dk/services/TMHMM

TOPCONS (consensus topology prediction membrane proteins): topcons.net

Index

| AAA+ proteins 245, 251–252, 254, 275, 278–280, 286, 392, 405–416 ABC transporters see transporters Aβ protein 59 Abri 59 accessible surface area 51 actin 48, 53–54, 62, 252–253, 393, 481–502, 504, 511, 561 barbed end 482, 484, 488, 490–491, 494 branching and crosslinking proteins see Arp2/3 complex capping proteins 482 fiber formation see formin and spire F (filamentous) actin 54, 481–483, 495 G (globular) actin 48, 54, 481, 483 pointed end 482, 488, 490, 496 severing proteins see gelsolin active site 48, 50–51, 140, 152, 154–156, 227–233, 236–240, 248–254, 260–265, 277, 283, 289, 292–297, 300, 327–328, 331–334, 348, 340, 342–347, 354, 358, 389–390, 404–405, 409–419, 435, 457, 460–463, 467, 472, 479, 482, 570, 574 acetylation see protein modifications acetyltransferase in fatty acid synthase see fatty acid synthase histone modifications see histone modifications | acyl carrier protein <i>see</i> fatty acid synthase acyltransferase 195 adaptive immune system <i>see</i> immune system adaptor protein 63, 315, 407, 458 adenine 105, 107–115, 124, 143–147, 156–157, 295, 300, 326, 347, 361, 406, 548 adenosylcobalamin 233, 235 adenovirus 540–541 adenylate cyclase 468, 476–478 adenylate kinase 49 adenylyl cyclase <i>see</i> adenylate cyclase aerobic metabolism 7 aerobic respiration 295 aflatoxin B_1 164–165 aggregation 32–33, 48, 52–54, 59–62, 168, 182, 269, 388–392, 401–402, 533 Agre P 429 aldolase 49 alphavirus 544–546 all- α proteins 42–51, 73, 79–95, 559 all- β proteins 42–51, 73, 96–102, 427–428, 559 allostery 47, 62, 229, 233–239, 393, 401, 463, 559 α / β proteins 46, 49, 234, 405, 559 α -crystallin domain (ACD) 391–392 α -helix <i>see</i> helix α -hemolysin 428 |
|--|---|
| acid base catalysis 153, 229 | Alzheimer's disease 59, 206, 391 |
| active transport 86, 90, 425–426, 443–446 | amino acids 14–19 |
| | |

ATP-binding domain 354, 356, 393

amino acid sequences 37-40, 43, 50, 78, 81, ATP synthase (ATPase) 241, 243, 245–256, 219, 269, 321, 385, 401, 419, 468, 482, 562, 438 binding change mechanism 252 aminoacyl-tRNA synthetases (ligases) (aaRS) catalytic mechanism 252-254 353-361 F₁ 248-255 F_o 248–255 aa-AMP 353, 358 classes 354-356 structure 248-251 subclasses 354-356 ATP synthesis 246–247, 252–255 editing 359–361 Avery O 7, 107 tRNA recognition elements 356–359 A-minor motif see RNA (structural motifs) bacteria 3, 5, 69, 79, 90, 96–97, 101–102, 113, amphiphiles 168, 177-185 164, 166, 192, 213–214, 218, 220, 222, 224, 245, 247, 249, 251, 256, 260, 275, 279, amphipathic character 70–71, 99, 163 amphipathic helices see membrane proteins 283–284, 286–287, 298, 301, 308, 320, 325, amyloid 33, 59-62, 386 329, 342, 351–352, 361–362, 364, 366–369, amyloid- β (A β) 59 379–381, 388–389, 392, 396, 402–403, 407, amyloid diseases 59 410, 417, 427–428, 444, 446, 481, 521, amyloid-forming proteins 59-63 535–536, 548 Anfinsen C 38, 385, 388 bacterial chemotaxis 313 antibodies see IgG bacteriophage 39, 49, 109, 287-288, 292, anticodon see tRNA 318–319, 325, 541, 559 antigen-binding domain(s) 522 HK97 540-541 antigens 521–525, 527–528, 533 MS2 134, 540–541, 547–548 T4 39, 235–236, 287–288, 537, 548–550, antigenic peptides 405, 416, 527, 530–531 antiporters see transporters 559 antitrypsin see protease inhibitors bacteriorhodopsin 8, 71–72, 427, 433, 438–439, apolipoproteins 206, 208–209 BAR domains 199-200 apoptosis 223, 323, 418, 453 base excision repair (BER) see DNA repair aptamer 142–143 base pair aquaporin 52–53, 87, 91, 429–430 selectivity filter 426, 429–432 classification 125 archaebacteria 166, 244, 361 conformational parameters 109–114 Arf see G-proteins Hoogsteen (H) 124–128, 136, 151, 295–296 arginine finger 245, 253–254, 277, 465, 467 mismatches 123 Arp2/3 (actin related proteins 2 and 3) 458, non-Watson-Crick 123-126, 135-136, 304 485, 490–493 reverse Hoogsteen 124-126, 136, 149, ASF1 273-734 aspartate carbamoyltransferase (aspartate reverse Watson-Crick 125-126 transcarbamylase) reverse wobble 126 Watson-Crick (WC) 114-115, 122-126, 132, aspartyl proteases see proteases ataxins 58 134, 141, 150–151, 294, 297, 304, 307, 372-373 atherosclerosis 206 atomic force microscopy 202 wobble base pair see also tRNA-mRNA ATPases 81, 241, 245, 251, 253 interaction 125–126, 373

base quadruplets 127–128

| bases (in nucleic acids) see nucleosides | C1q see complement system |
|--|---|
| base triplets 127–128 | Ca ²⁺ -ATPases <i>see</i> transporters |
| codons of mRNA 352-353 | cadherin see CAM |
| RecA 303-304 | calcium-gated potassium channel see ion |
| basic-helix-loop-helix (bHLH) proteins | channels |
| 317–318 | calmodulin 497 |
| Bawden FC 8 | calorimetry 169, 189, 202 |
| Bence-Jones proteins 523 | chameleon sequences 565 |
| Bernal JD 8 | cAMP 427, 433, 468, 478 |
| Berzelius JJ 7 | capping proteins see actin |
| β-adrenergic receptor see G-protein coupled | carbohydrates 3, 7–9, 213–224 |
| receptors | monosaccharides 213–216, 219, 221 |
| βαβ units 46 | deoxyhexose 215 |
| β-bulge 33–34, 101 | hexosamine 215, 221 |
| β-2 microglobulin 59, 528–529 | hexose 215 |
| β hairpin 43, 321, 455 | galactose 215 |
| β-helix or solenoid 43 | glucose 215–218, 220–222, 322, |
| β-sheets | 425–426, 444, 483 |
| antiparallel 32–33, 43–44, 325, 337, 355, | mannose 215, 220 |
| 377–378, 458–459, 479, 491, 497, 522, | pentose 109, 215 |
| 547 | ribose 109–115, 121–124 |
| barrel or cylinder 43–44, 46, 49 | xylose 215 |
| Greek key 43–44, 97 | uronic acid 215, 221, 223 |
| Jellyroll 43–44, 48–49, 540–541, 546 | sialic acid 215 |
| meander 43–44, 97, 491 | oligosaccharides 213, 216, 219-220, 223, |
| parallel 43–44, 46, 49, 242, 257, 369, 462, | 427 |
| 483, 504 | non-reducing end 216 |
| propeller 43–44, 338, 513 | reducing end 216 |
| sandwich 43–44, 48–49, 391, 454–455, 458, | polysaccharides 163, 213, 221–224 |
| 510, 522–523 | carbonic anhydrase 49, 51, 63, 229–233 |
| twist 32–33, 59, 98 | active site 231 |
| up-and-down 43–44, 97–98 | sulfonamide inhibitors 231 |
| β-turn 33–34 | cardiolipin see glycerophospholipids |
| B-factor 419, 556–557 | carotenoids 162 |
| binding change mechanism see ATP synthase | Caspar D 8, 537–538 |
| biotin 142–143, 256 | catalytic mechanism 50-51, 153-155, 222, 229 |
| BLAST 564 | 233–234, 241, 252, 254, 290, 294, 342, 388 |
| blood coagulation 224, 404, 418, 455 | catecholamines |
| BLOSUM62 570 | CATH see fold databases |
| bluetongue virus | caveolae 205 |
| bond energies 12 | caveolin 205 |
| bond lengths 12 | CD155 see poliovirus receptor |
| Boyer PD 252 | CD3 528, 533 |
| bracket notation 141–142 | CD4 528, 531–532, 543 |
| bulge see β bulge or RNA (structural motifs) | CD8 528, 531–322 |

chemiosmotic theory 245, 247

chirality 15, 29

consensus sequence (sequon) 134, 141,

219–220, 308, 324, 330, 337, 354, 465

| cooperativity 399–400, 417 | succinate dehydrogenase 246 |
|--|--|
| negative 400 | Deisenhofer J 72 |
| positive 400 | denaturation see protein denaturation or |
| CorA see ion channels | nucleic acid denaturation |
| cotransporters 425 | dengue virus 545 |
| corneocytes 175 | dephosphorylation 435, 452, 461 |
| CORN rule 15, 25 | depolymerization 54, 490 |
| Coulomb's law 13 | detergent 69, 72, 76, 166, 198, 201, 208 |
| covalent bond 11-13, 20, 228-229, 273 | dextro 15 |
| Creutzfeldt-Jakob's disease 59 | diacylglycerol see glycerolipids |
| Crick FHC 105–109, 125, 148 | dielectric constant 13 |
| cro repressor see transcription (bacterial | differentiation 224, 307, 526-527 |
| repressors) | diffusion rate 426 |
| cross-β 59–61 | disorder |
| cross-link 40, 63, 136, 482, 490, 493–494 | in lipid membranes 171, 173, 181, |
| Crowfoot-Hodgkin D 8 | 202–204 |
| cryo-EM 8, 271–272, 280–281, 366, 400, 415, | in proteins 47–48, 59, 239, 321, 330, |
| 495, 499, 536, 545, 550 | 456–457, 471, 483, 487, 504, 540–544, |
| crystallization 7–9, 52, 69, 72, 151, 159, 181, | 556, 567–569 |
| 469, 483, | distance plot 554 |
| crystallography see X-ray crystallography | disulfide bond 13, 19, 240, 388-389, 455, 462, |
| CTP:phosphocholine cytidyltransferase (CCT) | 509, 523 |
| 197–198 | divergence see evolution |
| cubic phase 174, 179, 181, 191, 195 | DNA (deoxyribonucleic acid) |
| cyclic GMP phosphodiesterase (PDE6) | 30 nm fiber 269, 271 |
| 473–475 | A-DNA 106–107, 116 |
| cyclophilin 386–388 | B-DNA 106, 109, 119, 303–304 |
| cyclosporine A 386 | backbone angles 109–114 |
| cysteine-rich domain 455, 458, 461–462 | double helix 108-109, 114-120, 122, 272, |
| cytochrome bc ₁ complex 246 | 297 |
| cytochrome c oxidase 85, 246 | double-stranded (ds, duplex) 111, 114, 116 |
| cytosine 105–109, 112–113, 115, 121, 153, | 121–122, 148, 269, 273–274, 277, 280–281 |
| 295–297, 300, 313, 316, 326 | 284, 288–289, 291–292, 326–327, 333–334 |
| cytokine receptors 453, 455, 461 | 338, 342–343 |
| cytoskeleton 481–482, 489, 496, 514 | major groove 115-116, 316-325, 337-338, |
| cytosol 3, 13, 71, 74–75, 92–93, 95, 98, 101, 396, | 340 |
| 401, 435, 440, 528 | minor groove 115–116, 291, 294, 297, 321, 324–325, 337–338 |
| Dali 554, 560 | non-coding 137 |
| deacetylation 315 | syn/anti conformation 113–114, 157, 296 |
| dehydratase see fatty acid synthase | Z-DNA 109, 116 |
| dehydrogenases 46–47, 52–53, 246, 503 | DNA-binding domain (DBD) 311-312, |
| lactate dehydrogenase 47, 53 | 316–317, 321–324, 456 |
| malate dehydrogenase 47 | DNA-binding proteins 115, 302, 321 |
| NADH dehydrogenase 246 | DNA damage 301, 346 |

| DNA helicases see helicases | ectodomain 511, 513–515 |
|--|---|
| DnaB family see helicases | effector site 237–238, 240 |
| DnaJ see heat shock proteins | EF-Tu, EF-G etc. (elongation factors) see |
| DnaK see heat shock proteins | translation |
| DNA ligase 273, 275 | EF hand 458, 494–497 |
| DNA polymerase | EGF see epidermal growth factor |
| classes 287–288 | EGF receptor (EGFR) 462 |
| exonuclease domain 273, 289-295 | elastin 224 |
| fingers domain 290–292 | electrochemical (or ion) gradients 69, 245-246, |
| Klenow fragment 288–294 | 255–256, 425–426, 434 |
| palm domain 289–293, 295 | electron diffraction 8, 71 |
| primase 288, 290, 294 | electron microscopy 8, 71, 162, 206, 275, 364, |
| primer (DNA or RNA) 273, 276, 286-287, | 366, 445, 483, 498, 548, 555 |
| 289–294 | electron transfer 64-65, 81, 274 |
| reverse transcriptase (RT) 288, 299, 300, 347 | electrostatic interaction 13, 31, 76, 197, 229, 247 |
| thumb domain 290, 292 | elongation factors see translation |
| DNA repair 273, 276–277, 288, 295–297, 323 | endoplasmic reticulum (ER) 93–94, 194, 197, 220, 388, 390, 481, 528 |
| base excision repair (BER) 297 | endosome 543 |
| 8-oxoG-DNA glycosylase (OGG1) | enoyl reductase see fatty acid synthase |
| 296–297 | enthalpy 14, 40 |
| nucleotide excision repair (NER) 297 | entropy 37, 39–40, 205, 247 |
| DNA-RNA hybrid see RNA polymerase | epidermal growth factor (EGF) 219, 462–465, |
| DNA synthesis see replication | 513–514 |
| DNA template 108, 114, 146, 272–276, 285–286, | module 514 |
| 288–295, 298–301, 307, 326, 328, 332–335, | receptor 462–465 |
| 340, 342–345 | epigenetics 271, 308, 312, 316 |
| DNA topoisomerase 128, 277, 282–285 | epinephrine (adrenalin) 451, 471, 475–477 |
| single strand (Topo IA, Topo IB) | epinephrine receptor 451, 468–469, 475–478 |
| 282–283 | ERK (extracellular-regulated kinase) see |
| double strand (Topo IIA, Topo IIB) | protein kinases (Ser/Thr kinases) |
| 284–285 | essential light chain see myosin |
| docosahexaenoic fatty acid 164 | etioplasts 194 |
| domain swap 54, 61, 395 | eukaryotes 6, 13, 79, 90, 93, 96, 152, 166, 214, |
| double bond 12, 20, 164–166, 186–187, 205, | 218–220, 244, 269, 272–276, 280, 283–288, |
| 387 | 294, 301, 307–308, 316, 325, 327, 329, 331, |
| drugs 69, 170, 175–176, 224, 282, 386, 395, 446, | 351, 361–362, 367–369, 377, 379–380, 388, |
| 475, 503, 546 | 390–392, 401, 403, 407, 413, 417, 430, 436, |
| drug delivery 170 | 488, 521 |
| proteins as drug targets 69, 446, 503 | evolution 1–10, 15, 43, 48–51, 55, 90, 121, |
| DsbA, DsbB, DsbC, DsbD disulfide bond | 233–234, 287, 301, 351–353, 386, 482, 498, |
| formation proteins 388–389 | 528, 559, 572–573 |
| dynamics see protein dynamics | convergent 51, 233 |
| dynein 503 | divergent 234 |

| exonuclease domain see DNA polymerase | proteins 37–41, 52, 59, 93, 101–102, 199, |
|---|---|
| extracellular matrix (ECM) 213–214, 223–224, | 219–220, 385–411, 559, 562, 565, 571 |
| 507–508, 511 | spontaneous 385–403 |
| extracellular signals 452 | assisted 386, 403 |
| eye lens 391, 573 | RNA 125, 137, 147–151, 156, 159 |
| | fork loop see RNA polymerase |
| F ₁ see ATP synthase | formin 485–488 |
| Fab see IgG | FH1 485, 487–488 |
| F-actin see actin | FH2 485, 487–488 |
| familial amyloidotic polyneuropathy 59 | lasso, linker, knob, coiled-coil, post |
| familial British dementia 59 | 487 |
| fatty acids 163–165, 188, 190, 215, 260, 262, | Fos see transcription factors (gene-specific) |
| 264–265 | four-helix bundle 46, 318, 454 |
| in lipids 161, 163–165, 188, 190 | Franklin, R 7, 105–107 |
| fatty acid synthase (FAS) 197, 260–265 | frustration (membrane) 180-182 |
| Type I 260 | FSI (formin spire interaction) 487–488 |
| Type II 260 | FtsH see proteases (oligomeric ATP-dependent |
| ferredoxin 49, 66, 262 | proteases) |
| ferritin 53, 65 | furin-like see cysteine-rich domain |
| FeS clusters 65 | fusion peptide 195–197, 542–546 |
| FfH signal recognition protein | |
| fiber diffraction 59, 105, 483 | G-actin see actin |
| fibrin 224, 508–509 | γ turn 23 |
| fibrinogen 512 | GAP (GTPase-activating proteins) 257–259, |
| fibrinolysis 418 | 373, 458, 465, 467, 472, 484 |
| fibronectin 224, 454–455, 461, 508–512, | RasGAP 458, 467 |
| 516 | GAL4 see transcription factors (gene-specific) |
| modules in proteins 508–510 | gated channels 426–427, 430 |
| type I 509–510 | GCN4 see transcription factors (gene-specific) |
| type II 509–510 | G-domain see G-proteins |
| type III 509–510 | GEF (guanine nucleotide exchange factor) |
| filamentous actin see actin | proteins 257, 374, 378, 458, 465–466, 468, 471, |
| finnish-type familial amyloidosis 59 | 474 |
| FK506-binding proteins 386 | gel phase 171, 173, 175, 186, 203 |
| flavivirus 140, 536, 544–545 | gelsolin 59, 485, 488–490 |
| fluid mosaic membrane model 199–201 | gene |
| F _o see ATP synthase | deletion 301 |
| fold classification 559 | duplication 1, 42, 90, 429, 489, 573 |
| fold databases 559–560 | expression 9, 113, 118, 137, 219, 270, 273, |
| CATH 43-44, 559-560 | 298, 301, 307–308, 323, 391, 452, 465, 491, |
| SCOP 43, 559–560 | 526 |
| fold recognition 569–571 | fusion 42, 575 |
| folding | insertion 564 |
| funnel 38 | genetic code 4, 120, 126, 147–148, 246, |
| process 37–39, 49, 385–386, 403 | 351–352 |
| | |

| codons 125, 129-130, 146-147, 150, 352, | glycosyl hydrolases 216–218 |
|--|--|
| 362, 367–368, 372–374, 377, 379–380, | glycosyl transferases (GT) 216–219 |
| 547–548 | Gly-zipper 55 |
| degeneracy 120, 126, 147-148, 352, 391, 411 | G-nucleotide exchange factor see GEF |
| genetic exchange (crossing over) 117, 301 | Golgi complex 3, 194, 197, 219–220, 528 |
| GFP (green fluorescent protein) 97, 491 | GPCR see G protein-coupled receptors |
| Gibbs energy 179–180, 247 | GPI (glycosylphosphatidylinositol) anchor 73 |
| Gibbs phase rule 170 | G protein-coupled receptors (GPCR) 427, |
| globin fold 47, 49 | 467–471, 475–478 |
| glucagon 451 | adrenergic receptor 468–469, 475–478 |
| glucagon receptor see G protein-coupled | glucagon receptor 468 |
| receptors (GPCR) | glutamate receptors 468–469 |
| glucocorticoid receptor see transcription | olfactory receptors 468 |
| factors (gene-specific) | pathways 468, 470, 476, 478 |
| glucokinase 561 | rhodopsin 427, 468–475, 477 |
| glucomannokinase 561 | G-proteins 205, 241, 243, 257–259, 368–370, |
| glucose see carbohydrates | 377, 452, 465, 468, 470–472, 477, 484, 491, |
| glucose transporters 425 | 498, 500, 503, 512, 514 |
| glutamate receptors see G protein-coupled | GAP see GAP |
| receptors | G-domains 257–259, 471 |
| glutaredoxin 235 | GEF see GEF |
| glycerol backbone 163–164, 166 | GTP hydrolysis 257–260, 371–374, 465, 467 |
| glycerol kinase 561 | Ras superfamily of monomeric G-proteins |
| glycerolipids 163–165 | 452 |
| diacylglycerol 195, 197 | Arf 452 |
| diglucosyldiacylglycerol (DGlcDAG) | Rab 452 |
| 189–190 | Ran 452 |
| monoglucosyldiacylglycerol (MGlcDAG) | Ras 49, 243, 257–258, 368, 370–371, |
| 174, 189–190, 192 | 452, 464–467, 471–472, 498, 554 |
| glycerophospholipids (phospholipids) | Rho 452 |
| 163–166, 168, 173, 202–203, 207–209 | trGTPases 363, 368–369, 373, 377 |
| cardiolipin 166 | trimeric G-proteins 243, 257, 370, 452, 458, |
| diphosphatidylglycerol 166, 188 | 468, 470–473, 476–478, 491, 512, 514 |
| phosphatidylcholine (PC, lecithin) 70, 75, | Gα 471, 473, 514 |
| 166, 168, 171, 179–180, 194, 197 | Gβ γ 470–471 |
| phosphatidylethanolamine (PE) 13, 75, | transducin |
| 188, 195–197 | grana stacks 194 |
| phosphatidylglycerol (PG) 75, 166, 188, 190 | Grb2 adaptor protein 458, 465 |
| phosphatidylinositol 166 | Greek key motifs see β -sheet |
| glycolipids 75, 167, 214, 218, 530 | green fluorescent protein see GFP |
| glycoproteins 213–214, 218–221, 224, 437, 526, | GroEL/GroES (GroE) 52–53, 391, 396–403 |
| 530, 536 | bullet model 398–400 |
| glycosylation 90, 218–220 | football model 399–400 |
| O-glycosylation 219 | refolding of misfolded proteins 391–392, |
| N-glycosylation 220 | 405 |

| growth hormone (GH) 451, 453–455 | helical net diagrams 57–58 |
|---|--|
| growth hormone receptor 453–454 | helical wheel diagrams 56–57, 78 |
| GrpE 391, 393, 561 | helicases 243, 245, 273, 275–287, 335 |
| G-tetrads see RNA (structural motifs) | DNA helicases 243, 276–282 |
| GTP hydrolysis <i>see</i> G-proteins | hexameric 243, 275, 278, 280–281 |
| GTPases see G-proteins | SF3, SF4, SF5, SF6 276–282 |
| guanine 105, 107–109, 112–115, 124, 126–129, | monomeric |
| 146, 154, 242, 295–296, 300, 324–325, 347 | SF1, SF2 277 |
| GxxxxGKT/S sequence motif (P-loop) | RNA helicases 276 |
| 242–243, 252–254, 257–259, 276–277, 301, | DEAD-box helicases 276 |
| 369–370, 374–379, 408, 467, 472, 498, 503 | helicase loader 243, 273, 275–276, 279–280, 286 |
| GxxxG sequence motif 83-84, 462 | helix-helix interactions 81–84 |
| • | helix-loop-helix motif 291, 317–319 |
| H ⁺ -ATPase <i>see</i> transporters | helix-turn-helix motif 284, 316–321, 330 |
| H, K ⁺ -ATPase <i>see</i> transporters | hemagglutinin see influenza virus |
| Haldane JBS 6 | heme 40, 63, 65 |
| halophilic 166 | hemerythrin 49, 53 |
| hammerhead ribozyme 138–139, 153–157 | hemicellulose 214, 222, 224 |
| handedness 29–30, 46, 56, 81, 97–98 | hemoglobin 7, 8, 42, 49 |
| left 24–25, 29–31, 56, 81, 83, 98, 109, 116, | Henderson R 8, 71 |
| 248, 269, 281, 303, 309, 459, 530 | heparan sulfate 221, 223–224, 508 |
| right 24, 29–31, 46, 55–56, 81–82, 97–98, | heparin 221, 223–224, 509–510 |
| 109, 115, 243, 279–281, 286, 303 | heptad repeat 55–57, 78, 317, 494 |
| haptens 524 | hereditary systemic amyloidosis 59 |
| heat shock proteins <i>see also</i> chaperones | herpes simplex virus 546 |
| HslVU see protease | hexagonal phase 171, 174, 180 |
| Hsp10 see GroEL/GroES | normal (H _I) 171, 174, 177 |
| Hsp40 (DnaJ) 391–395, 397, 402–403 | reversed (H _{II}) 171, 174, 177, 179–182, 188, |
| Hsp60 see GroEL/GroES | 193, 195, 197 |
| Hsp70 (DnaK) 391–397, 402–403, 408, 482, | hexokinase 482–483, 561 |
| 561 | Hidden Markov Model 571, 573 |
| Hsp90 391–392, 395–397 | histone acetyltransferases (HAT) 314 |
| Hsp100 (ClpA) see proteases | histone deacetylases (HDACs) 314 |
| Hsp104 (ClpB) see proteases | histone demethylases (KDM) 313 |
| Hsp110 391–392 | histone methyltransferases (HMT) 313 |
| small heat shock proteins (sHsps) | histones |
| 390–393 | H1 271, 316, 413 |
| helix | H2A 269, 273 |
| α-helix 24, 27–30, 46, 55, 57, 78, 80, 84, 206, | H2A.Z 310, 312 |
| 292, 316, 322, 453, 471, 517, 558, 563 | H2B 269, 273 |
| 3 ₁₀ helix 30 | H3 269–270, 273–274, 315, 338–339, 413 |
| helical bundle 65, 81, 90, 345, 407 | H4 269–270, 273–274, 314, 338–339, 413 |
| helical dipoles 30–31, 86–87, 441 | H5 270–271, 316 |
| π helix 30, 84 | histone code 271, 316 |
| poly-Pro helices 31, 459–460, 530 | histone fold 269-270, 338-339 |
| | |

| histone modifications 271, 273, 311–315 | 357–359, 370–374, 376, 388, 410, 420, |
|---|---|
| arginine methylation 313, 315 | 429–430, 472, 508, 516, 567 |
| lysine acetylation 313–315 | hydrogenase (Ni-Fe hydrogenase) 64 |
| erasers312, 315 | hydrophilic character 41, 65, 161, 168, 182, 220, |
| lysine metylation 313 | 398, 410, 420 |
| readers 312 | hydrophobic character |
| Bromo 315 | core 13, 37, 39, 42, 50, 56–57, 73, 75, 86–87, |
| PHD 315 | 98, 143, 566 |
| writers 312, 315 | effect 37–41, 52, 168, 182 |
| N-terminal tails 269–270 | (non-polar) interactions 14, 18, 37, 76–78, |
| octamer 269–270, 309, 334, 338 | 80–81, 83, 161 |
| HIV (human immunodeficiency virus) 132, | surface 389, 407, 430, 463 |
| 135, 137, 213, 387–388, 524–526, 536, 543, | hydroxyproline 31, 508, 512 |
| 559 | hypertension 206 |
| Env (=gp120 + gp41)525, 543 | Hünefeld FL 7 |
| gp120 525–526, 532, 543 | |
| HIV reverse transcriptase see DNA | I domain see integrins |
| polymerase (reverse transcriptase) | ICAMs (intracellular adhesion molecules) 512 |
| Rev protein 135 | icosahedral symmetry see symmetry |
| HLA see MHC | icosahedral viruses 54, 538, 540 |
| HNF-3 see transcription factors (gene-specific) | IgA (immunoglobulin type A) 522 |
| Hodgkin DC see Crowfoot Hodgkin | IgD (immunoglobulin type D) 522 |
| Holliday junction 117, 561 | IgE (immunoglobulin type E) 522 |
| homeobox sequences 320 | Ig fold (immunoglobulin fold) 48, 454, 458, |
| homeodomains see transcription factors | 511, 522, 533 |
| (gene-specific) | IgG (immunoglobulin type G) 522–523 |
| homologous proteins 49–50, 569, 572–573 | broadly neutralizing antibodies (bnAbs) |
| homologous recombination (HR) 297, 301, | 524–526 |
| 305 | CDRs (complementarity-determining |
| presynaptic filament 302–304 | regions) 522–524, 527, 530 |
| homology modeling 324, 569–571 | constant domain 522–523 |
| hormones 161–162, 164, 166, 322, 453 | Fab 522–524, 530 |
| HSlU see protease | Fc 522–523 |
| HSIV see protease | neutralizing antibodies (nAbs) 524–524 |
| Hsp proteins see heat shock proteins | variable domain 522–524, 530–531 |
| Huber R 73 | IgM (immunoglobulin type M) 522, 527 |
| humoral immunity 521 | immune system 48, 428, 451, 455, 521–533 |
| huntingtin 59 | innate 428, 521 |
| Huntington's desease 59 | non-adaptive 521 |
| hyaluronan 213–214, 221, 223–224 | antibody-mediated 521–527 |
| hydrogen bond 12–15, 18, 27–34, 37–41, 52, | T-cell mediated 527–533 |
| 57, 61–62, 74, 81, 83, 85–87, 97, 100, 107–108, | immuno-proteasome 416 |
| 114–115, 122–127, 133–136, 145, 150–155, | inclusion bodies 386, 391 |
| 157, 159, 168, 222, 231–232, 239–240, 257, | inflammation 418 |
| 270, 292, 294, 304, 312, 314–315, 324–325, | influenza virus 52, 213, 524, 536, 543–545 |

| hemagglutinin 52, 196, 536, 543, 545 neuraminidase 52, 524, 536 | kinesins C-, M- (or KinI) and N-type 498, 503–505 |
|--|--|
| initiation factors see translation | ncd (non-claret disjunctional) 505 |
| initiator tRNA see tRNA | kissing loops 132–133 |
| in-line S _N 2 attack 259–260 | Klenow fragment see DNA polymerase |
| inosine 112, 130, 148 | Klug A 8, 537–538 |
| inositol 166 | Kornberg A 288 |
| insulin 39, 572 | Kornberg R 330 |
| | ~ |
| insulin receptor 458, 461–462, 464 | Kv1.2 <i>see</i> ion channels (potassium channels) |
| integral membrane proteins see membrane | Citatitieis) |
| proteins | la stata dahardua annaga saa dahardua annaga |
| integrase 561 | lactate dehydrogenase <i>see</i> dehydrogenases |
| integrins see CAM | lamellar phase 174–176, 194, 202–204 |
| intercalation 337 | laminins 512 |
| interfacial region 70–71, 76, 87–89, 99, 168, | lateral pressure 182–185 |
| 566–567 | lectins 213 |
| interfacial tension 182, 184 | leucine-rich repeats 462 |
| interferons 451, 453 | leucine zipper 196, 317–319 |
| introns see mRNA | LeuT transporter 90, 440–441, 446–449 |
| ion channels | lever rule 173 |
| cGMP-gated sodium channels 473–474 | Levinthal C 38, 385 |
| Mg ²⁺ channel CorA | levo 15 |
| potassium channels 430–433 | light-chain amyloidosis 59 |
| KcsA 430–433 | light-harvesting complex 52 |
| voltage gated K ⁺ channels (Kv1.2) | lipids |
| 431–433 | aggregate structures 161–162, 168, |
| MthK 433 | 177–187, 195 |
| selectivity filter 426, 429–432 | categories 163–164 |
| ionotropic receptors see transporters | lateral diffusion 185–187, 189, 203–205 |
| islet amyloid polypeptide (IAPP) 59–60 | rafts 173, 186, 199–206 |
| isosteric 126, 149 | synthase 197 |
| | vesicles 77, 176–177, 179, 193–197, 199, |
| JAK Janus kinase family see protein kinases | 202, 481, 496, 502–503, 528 |
| (tyrosine kinases) | large unilamellar (LUV) 176–177 |
| jellyroll fold see β-sheet | small unilamellar (SUV) 176 |
| Jun see transcription factors (gene-specific) | lipoproteins 3, 206–208 |
| | LDL 206–208 |
| K ⁺ channels see ion channels (potassium | HDL 208 |
| channels) | liposome 162, 176 |
| KcsA see ion channels (potassium channels) | lyotropic liquid crystals 169 |
| keratin 42 | lyotropic polymorphism 169–170 |
| ketoacyl reductase see fatty acid synthase | liquid crystalline phases 168–170, 173–174, |
| ketoacyl synthase see fatty acid synthase | 178–179, 188 |
| KIM (kinase interaction motif) 465 | liquid-disordered (l _d) 204–205 |
| kinase domains 455, 459, 461–463 | liquid-ordered (l _o) 204–205 |

| LonA/B see proteases | multiple pass <i>or</i> polytopic |
|--|--|
| LUV see vesicles | 74–75, 79, 81, 84, 90–91, |
| lysidine 129–130 | 94–95 |
| lysine oxidase 508 | inverted repeat 90–91, |
| lysosomes 3, 404 | 429, 440–441 |
| lysozyme 39, 42, 59, 222, 524, 549–550, 559, | non-inverted repeat 91 |
| 573 | single pass <i>or</i> bitopic 73–75, |
| | 79, 81, 83, 90–95 |
| MacLeod C 7 | β-barrel <i>or</i> outer membrane |
| mad cow disease 61 | proteins (OMPs) 96–102 |
| major groove see DNA or RNA | peripheral membrane proteins 72–74, |
| major histocompatibility complex | 75–78, 89, 102, 194, 199 |
| see MHC | amphipathic helices 76–78, |
| malonyl/palmitoyl transferase see fatty acid | 197–198, 491 |
| synthase | conditional 73, 76 |
| Mat a1 <i>see</i> transcription factors (gene-specific) | monotopic 73 |
| Mat α 2 see transcription factors (gene-specific) | Meselson M 272 |
| MAP kinase <i>see</i> protein kinases (Ser/Thr | messenger RNA see mRNA |
| kinases) | metabolism 1–2, 5, 7, 205, 219, 272, 315, |
| maturation cleavage 546 | 322 |
| Max see transcription factors (gene-specific) | metabotropic glutamate receptor family see |
| M-band see muscle | G protein-coupled receptors |
| McCarty M 7 | metabotropic receptors 427 |
| mechanosensitive channels 185, 433 | metal binding 63–66, 514–515 |
| Mediator see transcription factors (general) | metal clusters 65 |
| medin 59 | metal ion dependent adhesion site (MIDAS) |
| meiosis 301 | see integrins |
| membrane(s) | 9 |
| | metastable proteins 418 |
| biogenesis 197 | methylated bases see nucleosides |
| curvature 77, 179, 195, 198–200 | methylation see protein modifications |
| mean 204 | methyl transferases 313 |
| monolayer 177, 179–181, 185–186 | Mg ²⁺ channel <i>see</i> ion channels |
| spontaneous 177, 179–180, 194–195, | MHC (major histocompatibility complex) or |
| 198 | HLA (human leucocyte antigen) 405, 416, |
| total 179 | 521, 527–533 |
| domain formation 176, 188, 201–202, | class I 416, 527–532 |
| 204–205 | class II 527–532 |
| fission 194–196 | Michel H 72 |
| fusion 176–177, 194–197, 525, 542–546 | Miescher F 7 |
| stalk formation 195 | microfilaments 481, 491, 502 |
| membrane attack complex 428 | microtubules 54, 502–505 |
| membrane proteins | MIDAS motif see integrins |
| integral <i>or</i> transmembrane proteins (TMP) | Miller S 6 |
| 57, 73, 79–96, 566 | minor groove see DNA or RNA |
| α-helical 79–84, 90 | miRNA (microRNA) 307 |

| misfolded proteins 400, 405 | mutations 19, 39–40, 109, 120, 122, 130, 132, |
|---|---|
| Mitchell P 246–247 | 149, 157, 273, 294–295, 323, 346, 379, 390, |
| mitochondria 13, 96, 98, 101, 161, 166, 245, | 417, 441, 467, 527, 573 |
| 247, 251, 297, 352, 361–362, 396, 410, 427, | Myc see transcription factors (gene-specific) |
| 502 | MyoD see transcription factors (gene-specific) |
| mitochondrial membrane 194, 248 | myoglobin 8, 30, 40, 42 |
| molecular dynamics 324, 431, 441, 571 | myosin I, II, V, VI 496–502 |
| molecular evolution 48–51, 55, 233–234, 287, | actin binding 493–495, 501, 504 |
| 301, 351–353, 559, 572–573 | conformational states 500–502 |
| molecular genetics 301 | rigor, post-rigor, pre-power stroke, |
| molecular motors (see also myosin or kinesin) | power stroke 500–502 |
| 229, 276 | converter domain 497, 500 |
| molecular switches see G-proteins | essential light chain (ELC) 497, 501 |
| molten globule 39, 419 | heavy chain 496–497, 501 |
| monoamine oxidase 79 | regulatory light chain (RLC) 497, 501 |
| monolayer curvature see membrane(s) | S1 fragment 497, 499 |
| motility 481–505 | similarities with G-proteins 498 |
| motor domains, motor proteins see myosin or | myristoylation see protein modifications (lipid |
| kinesin | modifications) |
| mRNA 105, 122, 128, 141, 146–147, 152, 156, | |
| 316, 330, 344, 346–347, 351–352, 361–363, | Na ⁺ ,K ⁺ -ATPase <i>see</i> transporters |
| 366, 368–369, 372–373, 377, 379–381, 502, | Na ⁺ /H ⁺ antiporter 440 |
| 526 | nascent polypeptide see translation |
| coding region 121, 137, 146, 308 | natively unfolded/unstructured protein 48 |
| codon see genetic code | ncd see kinesins |
| codon-anticodon interaction see | neural network 563–565, 568 |
| translation | neuraminidase see influenza virus |
| initiation codon (AUG) 352, 367, 547-548 | neurotransmitters 446, 474–475 |
| introns 139, 152, 156–157, 309, 330, | neutron diffraction 41 |
| 346–348 | NhaA see transporters |
| group I intron 152, 156–157 | Nicolson GL 199, 201 |
| group II intron 152 | nitrogen fixation 64 |
| poly(A) tail 146 | NMR spectroscopy 9, 41, 189, 208–209, |
| stop codon 146-147, 352, 362, 379-380 | 555 |
| UTR (untranslated region) 140, 146 | non-Watson-Crick base pairing see base pair |
| MthK see ion channels (potassium channels) | non-coding DNA see DNA |
| multidomain proteins 257, 260, 458 | non-histone chromosomal proteins 269 |
| multifunctional enzymes 260 | nonlamellar phases 173-174, 188, 190-191, |
| muscle 54, 205, 318, 434, 481–482, 493–501 | 194 |
| contraction 493–494, 498–502 | Northrop JH 8 |
| M-band 493–494 | NSS see transporters |
| sarcomere 493–494, 501 | Ntn-hydrolases 416–417 |
| thick filaments 493-494, 496, 501 | nuclear receptors 317, 322 |
| thin filaments 54, 493-494, 501 | nuclease 38, 271, 385 |
| Z-disc 493–494, 501 | nucleic acid denaturation 118, 276 |

| nucleic acid structure see DNA or RNA | p23/Sba1 co-chaperone <i>see</i> chaperones |
|---|--|
| nucleocapsid 542 | P4-P6 domain see ribozymes |
| nucleophilic attack 154, 156, 231–232, 347, | p53 see transcription factors (gene-specific) |
| 435 | packing parameter 177–179, 195 |
| nucleosides 109–110 | palindromic sequences 319, 321, 325 |
| bases 112-113 | palm domain see DNA polymerase |
| methylated bases 113 | palmitoyllinoleoyl -PC 166 |
| syn/anti conformation 113–114, 157, 296 | palmitoylarachidonyl -PC 166 |
| nucleoside analogs | palmitoyldocosahexaenoyl-PC 166 |
| AZT | Parkinson's disease 59, 391, 446 |
| nucleosome 269–271, 309–313, 316, 338 | partial charges 19, 31 |
| histones see histones | passive transport 248, 425–427 |
| linker DNA 269, 271, 309 | pathogen 527, 530 |
| non-histone proteins 313 | PDB see Protein Data Bank |
| nucleosome-depleted regions (NDR) 310 | PDI see protein disulfide isomerase |
| nucleosome-free regions (NFR) 310 | pepsin see proteases |
| tetranucleosomes 271–272 | peptide bond 20–23 |
| nucleotide exchange factor (see also GEF | cis 20, 22, 31 |
| proteins) 257, 369, 374, 378, 391, 465, 488, | trans 20, 22, 27, 31 |
| 561 | peptidoglycan (murein) 213–214, 224 |
| nucleotide excision repair (NER) see DNA | peptidyl-prolyl-cis/trans isomerase (PPIase) |
| repair | 402 |
| nucleotides 109–110 | peptidyl transfer <i>see</i> translation |
| nucleus 3–4, 297, 308, 323, 330, 451–453, 456, | peripheral membrane proteins <i>see</i> membrane |
| 464–465, 488, 542 | proteins |
| 101 100/ 100/ 012 | Pfam database 573–574 |
| OB-fold 415 | phagocytosis 521–522 |
| Oct-1 see transcription factors (gene-specific) | phase diagram 169–175, 181, 202–204 |
| odorant receptors see G protein-coupled | ternary 172–173, 203–205 |
| receptors (olfactory receptors) | three-phase line 173 |
| odorants 451 | phase transition 169–170, 173, 175–176, 182, |
| Okazaki fragments see replication | 189, 191–193, 203 |
| olfactory receptors see G protein-coupled | PHD 315, 564 |
| receptors | PHDsec 564–565 |
| ω (omega) torsion angle 22 | phosphatases 452, 457–458, 465 |
| OmpF 72, 97, 99 | Phosphatidylcholine 70, 166, 168, 179, |
| Oparin A 6 | 194 |
| O-phosphatidyltrimethylarsonium lactic acid | phosphatidylinositol <i>see</i> glycerophospholipids |
| 168 | (phospholipids) |
| opsin see G protein-coupled receptors | phospholipases 166–168, 194–195, 458, 468 |
| (rhodopsin) | phospholipase C 167, 194, 458, 468 |
| ovalbumin 418 | phospholipids see glycerophospholipids |
| Overton EC 161 | phosphonolipid 167 |
| | |
| oxidative phosphorylation 246–247 | phosphopanteine transferase <i>see</i> fatty acid |
| oxoG glycosylase 296–297 | synthase |

| phosphorylation 215, 246–248, 312, 391, 435, | profile methods 569 |
|--|--|
| 451–452, 457, 459–460, 463, 527 | profilin 484–485, 487–489 |
| photosynthesis 7, 64, 386 | ProFunc 574 |
| photosynthetic reaction center 72, 88 | promoter 128, 308–310, 323, 325–326, 329, 330, |
| phytanyl 166 | 334–342 |
| π-electrons 19–20 | propeller protein 43, 119, 256, 338, 458, 461, |
| Pirie NW 8 | 470, 491, 512–517 |
| pleckstrin (PH) homology domain 458–459 | ProSite 573–574 |
| P-loop motif see GXXXXGKT/S | prostaglandins 164 |
| pol I, pol II, pol III see RNA polymerase | prosthetic groups 63–66 |
| (eukaryotic) | proteases (proteolytic enzymes) 42, 50–51, |
| poliovirus 536, 546–547 | 223–224, 229, 329, 385, 404–407, 409–411, |
| poliovirus receptor 546 | 413, 522, 543, 559 |
| poly(A) tail see mRNA | aspartyl (acidic) proteases 42 |
| polyadenylate polymerase 146 | carboxypeptidase 404 |
| polyamines 151, 433 | oligomeric ATP-dependent proteases |
| polyketides 163–165 | 405 |
| polyomavirus 538 | ClpAP/ClpXP 405 |
| polypeptide exit channel <i>see</i> ribosome | FtsH 243, 245, 405, 410–411 |
| polyproline helix 31, 459–460, 530 | HslUV (ClpQ) 405, 407-410 |
| porins 96, 427–428 | Hsp100 243, 391–393, 405, 407–409 |
| OmpG 427–428 | LonA/B 405 |
| α-hemolysin 428 | proteasome see proteasome |
| posttranslational modifications 323 | unfolding 403–406 |
| positive-inside rule 89–90, 95, 443, 566, | Ntn-hydrolases 416–417 |
| 568 | papain 404, 523 |
| potassium channels see ion channels | pepsin 8, 404 |
| PPIases (proline <i>cis-trans</i> isomerases) 386–387, | serine proteases 50–51, 229, 522 |
| 395, 402 | catalytic triad 50–51 |
| cyclophilin (CypA) 386–388 | chymotrypsin 3, 50–51, 404, 417, |
| FKBP 386–387, 395 | 566 |
| parvulin 386–387 | subtilisin 49–51, 404 |
| prediction of | trypsin 229, 404, 417 |
| disordered segments 567-569 | protease inhibitors 418–421 |
| function 574 | antichymotrypsin 481, 421 |
| secondary structure 562–569 | antithrombin 224, 418 |
| tertiary structure 569–571 | antitrypsin 418–419, 421 |
| topology of transmembrane proteins | ovalbumin 418 |
| 566–567 | plasminogen activator inhibitor 418 |
| preinitiation complex see transcription | serine protease inhibitors (serpins) 224, |
| prenol lipids 164–165 | 418–421 |
| Pribnow box 308 | reactive center loop (RCL) 418-421 |
| primary transport 425, 433–434 | proteasome 26S 405, 413–416 |
| primase RNA polymerase 276, 286 | 11S activator complex 416–417 |
| prion protein 59, 61, 418 | 19S regulator complex 413–416 |

immuno-proteasome 416 proton motive force 246–247, 252, 254, 444 11S/PA26 complex 416 PSI-BLAST 569 proteolytic chamber 20S 407, 409, 413-417 pucker see sugar pucker pulmonary surfactant membrane 203 unfolding 403–406 Protein Data Bank (PDB) 25, 42, 63, 69, 555, pulsed field gradient NMR 186, 204 purine bases see nucleosides 558-559 purple membrane 71 protein denaturation 40, 118, 386, 390 protein disulfide isomerase (PDI) 386, 388-390 pyrimidine bases see nucleosides DsbA, DsbB, DsbC, DsbD 388–389 pyrrolysine 15 protein dynamics 208 protein folding see folding process quadruplex see RNA (structural motifs) protein kinases 194, 395, 452, 456–458, 462, 464 quasi-equivalence 537–538 Ser/Thr kinases 452, 457 MAP kinase (ERK) 462, 464–465 Rab see G-proteins tyrosine kinases see also receptor tyrosine Rad51 recombination proteins 244, 301–302 kinases (RTKs) 455, 457-464 radicals 233, 295 cyclin-dependent kinase 457, 463 radical generator see ribonucleotide reductase JAK kinases 453, 455–456, 462 Raf (MapKKK) 464–465 src kinase 458–460, 533 Ramachandran plot 23–25, 27, 31, 555, 557 kinase C 194, 458, 460 Ran see G-proteins Ras see G-proteins protein kinase domain 455–456, 459–464 protein misfolding disorders (PMDs) 59, 61 RasGAP see GAP reaction center see photosynthetic reaction protein modifications acetylation 312–315 glycosylation 90, 218–220 RecA 243–244, 276, 278–280, 301–305 receptor tyrosine kinases (RTKs) 453, 455, hydroxylation 18 methylation 312-313, 315 461-467 phosphorylation 312, 391, 435, 451–452, epidermal growth factor (EGF) receptor 457, 459–461, 463, 527 462-464 SUMOylation 312, 314 insulin receptor 458, 461-462, 464 ubiquitination 312, 314 recombination 244, 272, 275, 297, 299, 301–305 oxidation 388–390 RecQ helicase see helicases protein phosphatases 452, 457–458 regulatory light chain (RLC) see myosin protein secondary structure 28-35, 562-569 replication (DNA synthesis) 3, 5–7, 105, 113, propensity 29, 562-563 118, 243, 269, 272–301, 535 protein synthesis see translation accuracy 287-288 protein tertiary structure 35, 37, 42–51, Cdc6 273, 275, 280-281 569-571 clamp loader 243, 273, 285-287 protein turnover 404 CMG complex 280–281 protechuate 3,4-dioxygenase 53 direction of synthesis 274-276, 280-281 proteoglycans 214, 218, 221, 223–224, 507 lagging strand 275, 280-281, 286-288, 294, proteolysis 219, 403–404, 418–419, 507 proteolytic chamber see proteasome leading strand 274, 280–281, 286, 288, protofilament 502 proton gradient 246-249, 252, 438-440 MCM complex 273, 276, 280-281

| Okazaki fragments 275, 286, 298 | 23S rRNA 132, 136, 139, 141, 144–145, |
|---|--|
| origin of replication 243, 275–276, 279–281, | 361–362, 365, 373–377, 379 |
| 286, 298 | 28S rRNA 361, 362 |
| origin recognition complex (ORC) 273, | 5S rRNA 134, 136–137, 329, 361–362, |
| 275, 280–281 | 365, 367 |
| primer 273, 276, 286–287, 289–294, | 5.8S rRNA 361–362 |
| 299 | A1493, A1492, G530 372–373 |
| replication bubble 282 | sarcin-ricin loop (SRL) 133, 136, 147, 373 |
| replication fork 273–277 | A2662 373–374 |
| replication complex (replisome) 276, 278, | ribosome |
| 280–281, 286 | 30S subunit 141, 361–363, 368 |
| reverse transcription see DNA polymerase | 40S subunit 361–362, 366 |
| sliding clamp 273, 276, 285–287, 295 | 50S subunit 141, 143–144, 159, 361–363, |
| repressors see transcription | 366 |
| reproduction 1–2 | 55S ribosomes (mammalian mitochondria) |
| resolvase 561 | 362, 367 |
| respiratory chain 246 | 60S subunit 361–362, 366 |
| retina 474 | 70S ribosomes (archaeal and bacterial) |
| retinal 438, 472–474 | 362–363, 382 |
| all <i>trans</i> 472–474 | 80S ribosomes (eukaryotic) 362, 366 |
| 11– <i>cis</i> 472–474 | A-site 344–345, 363, 368, 372, 374–379, |
| reverse transcriptase (RT) see DNA | 381 |
| polymerase | decoding site 363, 374, 380–381 |
| RF1, RF2, RF3 release factors see translation | E-site 368–369, 376–377, 382 |
| rhinovirus 546 | hybrid sites 364, 376, 378 |
| Rho see G-proteins | polypeptide exit channel 375–376, 402 |
| rhodopsin see G protein-coupled receptors | peptidyl transfer site 372, 379 |
| ribonucleases (RNases) 141, 146, 152, 277 | P-site 362, 367–369, 375–379, 382 |
| MRP 141 | ribosome-nascent chain complex (RNC) 92-93 |
| RNase P 152 | 102 |
| ribonucleotide reductase (RNR) 233-240 | ribosome recycling factor (RRF) 381-382 |
| active site 236–240 | ribozymes 122, 138–139, 152–159, 347, 375 |
| allosteric regulation 234, 236 | hammerhead ribozyme 138-139, 153-157 |
| overall activity site 237 | P4-P6 139, 156–159 |
| radical generator 235 | ricin 49, 147 |
| RNR classes 233, 235–237 | RNA (ribonucleic acid) |
| specificity site 233, 236–237, 239 | A form 122, 126, 135 |
| thiyl radical 233-234, 240-241 | deep groove (major groove) 122-123, 127, |
| ribosomal proteins 47, 136, 141, 147, 310, 351, | 151, 157 |
| 361, 365–367, 375, 378, 402 | double helix 121-122, 130-133, 137-144, |
| ribosomal RNA 122, 128, 130, 136, 138, 143, | 276, 558 |
| 147, 152, 307, 329, 351, 361, 366–367 | hairpin structure 122, 128–129, 132–134, |
| 16S rRNA 121, 131, 134, 139, 361–362, | 139, 152 |
| 366–368, 373 | internal loops 135–137 |
| 18S rRNA 361, 366 | junctions 138–140 |

| modified bases 128–130 | RNAse see ribonuclases |
|--|--|
| secondary structure 123, 130-132, 136-137, | RNA world 152, 233 |
| 141–143, 146–148, 153–154, 157–158 | RNR see ribonucleotide reductase |
| shallow groove (minor groove) 122, 135, | rod-shaped virus particles 537 |
| 139, 141, 144–145, 156–157 | Rossmann fold 46, 257, 264, 354, 411, 503, 512 |
| single stranded 111, 120-122, 130, 133, 148, | rotamers 25–27 |
| 547 | rotavirus 536 |
| structural motifs | Rous sarcoma virus 460, 561 |
| A-minor motif 143–145 | rRNA see ribosomal RNA |
| bulge 138–139 | rudder see RNA polymerase |
| G-tetrads 127–128 | RuvB see helicases |
| K-turn or kink-turn 141 | |
| pentaloops 132–134, 142 | saccharolipids 163–165 |
| pseudoknot 127, 142–143 | S-adenosyl methionine 63, 235, 313 |
| quadruplex (tetraplex) 127–129 | sarcin 147 |
| tetraloops 132–134, 156–159 | sarcin-ricin loop see ribosome |
| ANYA 134 | SAXS 169, 187 |
| GAAA 134, 156–159 | SCOP see fold databases |
| GNRA 133–134, 157 | SCOR database 134 |
| UNCG 134 | scoring matrices 570 |
| triplets 127–128 | secondary structure see protein secondary |
| U-turn 133–134, 150–151, 154 | structure or RNA (double helix) |
| tertiary structure 120–123, 127, 132, | secondary structure prediction 142, |
| 143–144, 148–149, 153, 159 | 562–569 |
| triple helix 127–128 | protein 562–569 |
| RNAi 307 | RNA 130–132, 141 |
| RNA polymerase 146, 219, 276, 282, 307–308, | secondary systemic amyloidosis 59 |
| 310, 325–346, 361, 479 | secondary transporters 425–426, 437, |
| active site 327–328, 331–334, 340, 342–347 | 439–449 |
| backtracking 326-327, 345-346 | secretin-like GPCR family see G protein- |
| bacterial 327–329 | coupled receptors |
| bridge helix 332–333, 343–345 | selectivity filter see ion channels (potassium |
| clamp 327, 332–333, 340, 342–343, 346 | channels) |
| DNA-RNA hybrid 326–328, 333, 343–344, | selenocysteine 15, 369 |
| 346 | senile systemic amyloidosis 59 |
| eukaryotic 326, 329–346 | sequence alignment 553-554, 564, 569-575 |
| pol I 329 | multiple (RNA) 130, 147 |
| pol II 329–346 | protein 563–564, 567 |
| pol III 329 | sequence families 560 |
| fork loops 333, 343 | sequence patterns 564, 572–573 |
| hybrid helix 326, 343, 345–346 | serine proteases see proteases |
| jaws 331–332, 342–343 | serine protease inhibitors see protease |
| RNA dependent 307 | inhibitors |
| trigger loop/helix 333, 344–346 | Ser/Thr kinases see protein kinases |
| wall 333, 340, 343 | serum amyloid A 59 |

| seven-transmembrane helix proteins see | Stahl F 272 |
|--|---|
| G protein-coupled receptors <i>or</i> | Stanley WM 8 |
| bacteriorhodopsin | stacking 20, 59, 115, 121–124, 132–135, 141, |
| SH2 domains 456, 458–463, 465 | 151, 154–157, 208, 239, 358–359 |
| SH3 domains 199, 458–461, 465 | STAT proteins (signal transducers and |
| Shc adaptor protein 458 | activators of transcription) 453, 456, |
| Shine-Dalgarno region 368 | 462 |
| sHsps see small heat shock proteins | stathmin 503 |
| σ factor see transcription factors (general) | steroid hormones 322 |
| signal recognition particle (SRP) 92–93, 139, | sterol lipids 163–165 |
| 401 | cholesterol (CHOL) 3, 70, 163, 172-173, |
| signal transduction 201, 257–258, 425, | 186–188, 201–208 |
| 451–479 | "bad" 208 |
| signaling pathways 63, 257, 451–453, 456, 459, | "good" 208 |
| 461–463, 467, 469–470, 473, 478 | stop codon see mRNA |
| immediate effects see GPCRs | strand-exchange 244, 301 |
| lasting changes see protein kinases | stratum corneum 175–176 |
| silk 42 | streptavidin 253 |
| Singer SJ 199, 201 | streptolydigin 344–345 |
| siRNA 122 | stress fibers 491 |
| sliding clamp see replication | structural alignment 553 |
| small heat shock proteins see heat shock | structural convergence 572 |
| proteins | structural genomics 574 |
| snake venoms 166 | structural superposition 553 |
| SNARE proteins 196–197 | subtilisin see proteases (serine proteases) |
| son of sevenless (Sos) 465–466 | suicide inhibitors 418 |
| southern cowpea mosaic virus 542 | sugar pucker 111, 117, 122–123, 133, 151, 558 |
| spermidine 151 | C2'-endo 111-112, 117, 123, 151 |
| spermine 151 | C3'-endo 111–112, 117, 122–123 |
| sphingolipids 163–165, 167, 197, 201, 204 | Sumner JB 8 |
| ceramides 197 | SUMOylation see protein modifications |
| glycosphingolipids 204 | superfamilies 43, 199, 242, 392, 409, 413, |
| cerebroside 193 | 439–441, 452, 559–560 |
| phosphosphingolipids 172, 182, 187, 201, | superoxide dismutase 19, 49 |
| 204 | superresolution fluorescence microscopy |
| spire 484–488 | 201 |
| FYVE domain 486, 488 | SV40 538 |
| KIND domain 486–488 | Svedberg T 8 |
| splicing 139, 152, 156, 323, 330, 346–348, 508 | SwissModel 570 |
| spliceosome 139, 346–348 | SwissProt 572–574 |
| U1, U2, U4, U5, U6 346–348 | Symmetry 8, 52–54, 65, 90–91, 107, 124, |
| sponge phase, L ₃ 181 | 126, 250–251, 254, 261–262, 278–279, |
| src kinase see protein kinases (tyrosine | 286, 316, 318–319, 338, 377, 410–411, |
| kinases) | 413, 415–416, 430, 441, 443, 447, 505, |
| S-S bond see disulfide bond | 535-540, 542, 544, 548 |

| icosahedral symmetry 53, 537–540, 544 | torsion angles 20, 23, 25–27, 33, 110–112, 114, 386, 555 |
|--|--|
| two-fold 90-91, 377, 411, 443, 505, | trans see configuration |
| 537 | transactivation domains 316, 323 |
| symporters see tranporters | transcription (RNA synthesis) 3–5, 7, 105, 110, 146, 219, 269–271, 273, 276, 282, 299, |
| T4 lysozyme see lysozyme | 307–347, 386, 412, 451, 453, 456, 464–465, |
| TAFs see transcription factors (general) | 526 |
| talin 515 | activators 307-308, 316-326, 330, 334-335, |
| Tanford C 168 | 337–338, 341 |
| Tat protein 137 | bacterial repressors 307–308, 316–325, |
| TATA-box see transcription | 452 |
| TATA box-binding proteins (TBP) see | arc 325 |
| transcription factors (general) | cro 318–320 |
| TBP-associated proteins (TAFs) see | lambda 318–319, 321 |
| transcription factors (general) | trp 317, 320 |
| T-cell receptors (TCR) 521, 528, | hybrid helix 326, 343, 345–346 |
| 530–533 | preinitiation complex 330, 333–336 |
| CDRs 522, 524, 527, 530 | recognition element 340 |
| telomerase 288, 298–301 | start site (TSS) 301, 307–308, 312, 333, 337, |
| telomerase RNA (TER) 299–300 | 339 |
| telomerase protein (TERT) 299–300 | TATA box 308, 334–335, 337–338, 340–341, |
| telomeres 128, 298–299 | 343 |
| ternary phase diagram 172–173, 203–204 | transcription bubble 326–327, 329, |
| TF (trigger factor) see chaperones | 333–335, 337, 340, 342–343 |
| TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, TFIIS | transcription factors (general) 335–341 |
| see transcription factors (general) | Mediator 341–342 |
| thermal denaturation 40, 118 | σ 329, 330, 335, 336 |
| thermophilic bacteria 40, 166 | TFIIA 330, 335–336 |
| thermosome 396 | TFIIB 330, 333, 337, 339–341 |
| thioredoxin 235, 388, 390 | TFIID 323, 330, 334–340 |
| threading methods 569 | TAFs 334–335, 337–338 |
| thrombin 224 | TAF1 338 |
| thylakoid membranes 90, 162, 194 | TBP (TATA box-binding protein) 330, |
| thymine 105, 107–109, 112–115, 121–122, | 334–335, 337–342 |
| 307 | TFIIE 330, 335–336, 340–341 |
| thymidine kinase <i>see</i> deoxyribonucleoside | TFIIF 329–330, 334–337, 339 |
| kinases | TFIIH 276, 323, 329, 335–336, |
| tymosin β4 48, 484–486 | 340–341 |
| TIM barrel 45–46, 48–49, 222, 265 | transcription factors (gene-specific) 308, 312, |
| titin 494 | 316–325 |
| $T_{\rm M}$ main transition temperature 170, 187, 203, | basic-leucine zipper (bZIP) proteins |
| 205 | 317 |
| tobacco mosaic virus (TMV) 53, 537 | GCN4 317–318 |
| topology diagrams 89–91 | MyoD 317–319 |
| T | , |

| Fos 317–318 | IF1 368 |
|---|--|
| GAL4 317 | IF2 368, 378 |
| HNF-3 | IF3 368, 382 |
| homeodomains 317, 320–321 | nascent polypeptide 90, 93, 365, |
| MAT a1 320–321 | 375–376, 401–402 |
| MAT α2 320–321 | peptidyl transfer 361–363, 365, 368, |
| Oct-1 321 | 372, 375–377, 379 |
| Jun 317–318 | proofreading 374 |
| MAX 317, 319 | release/termination factors |
| Myc 317–319, 325 | 379–381 |
| p53, p63, p73 323–324 | RF1, RF2, RF3 379-381 |
| pioneer 316–317 | ribosome recycling factor (RRF) |
| FoxA 316 | 381–382 |
| PU.1 316-317 | switch I and switch II 370-371, 374, |
| zinc fingers 321–322 | 378–379 |
| glucocorticoid receptor 322–323 | tetracycline resistance (TetM etc) |
| Zif268 317 | 369 |
| transdermal delivery 175 | translocation 346, 362, 364, 368–369, |
| transducin 468, 470–475 | 377–379 |
| transient interactions 9 | translocon 93–95, 101–102 |
| transition state 155–156, 228, 232–233, 244, 259, | transmembrane gradient 247 |
| 293–294, 467, 472–473 | transmissible spongiform encephalopathy |
| transition state analogs 231–232, 472 | (TSE) 61 |
| translation (protein synthesis) 351–382 | transporters |
| codon-anticodon interaction 368, | channel proteins 55, 81, 426, |
| 372–373 | 427–433 |
| A1492, A1493, G530 372–373 | gated or ionotropic receptors 427, |
| elongation factors 136, 147, 152, 242, | 430–433 |
| 362, 368–379 | metabotropic receptors (e g |
| EF-G, EF2 136, 242–243, 368–369, | GPCR) 427, 467–479 |
| 377–379, 382 | primary 425, 433–439 |
| EF-Ts, EF1B 369, 374–375, 378 | pumps 425–426, 429, 433–435, |
| EF-Tu, EF1A 136, 152, 243, | 438–439 |
| 362–363, 368–372, 374–375, | P-type ATPases 434–436 |
| 378–379, 381 | Ca ²⁺ -ATPases 434–436 |
| EF-Tu-EF-Ts complex 374 | H^{+}/K^{+} -ATPases 434–436 |
| LepA, EF4 369 | Na ⁺ , K ⁺ -ATPase 434–436 |
| fidelity 353, 373 | ABC transporters 436–438 |
| initial tRNA binding/selection 362, | secondary 425, 439–449 |
| 364, 368, 371–372 | antiporters 440–446 |
| initiation factors | NhaA 440–443 |
| eIF2 368 | small multidrug resistance |
| eIF4A 146 | (SMR; EmrE) 442–444 |
| eIF4B 146 | resistance-nodulation-division |
| eIF5B 368 | (RND; AcrB) 444–446 |

symporters 446–449

| neurotransmitter sodium | • |
|---|--|
| symporters (NSS; LeuT) | ubiquitin 49, 314, 405, 411–418 |
| 446–449 | iso-peptide linkages 412 |
| transthyretin (prealbumin) 53, 59, 61 | ubiquitin-activating protein (E1) 411–412 |
| triangulation number 538–540 | TAF1 338, 340 |
| tricyclic antidepressants 446, 449 | ubiquitination 312, 314 |
| trigger factor (TF) see chaperones | ubiquitin-conjugating enzyme (E2) 411–412 |
| triose phosphate isomerase (TIM) 46, 49 | ubiquitin-protein ligase (E3) 411–412 |
| tRNA 147–152 | ubiquinone 161 |
| acceptor arm/stem 148–149, 151, 352, | ultracentrifuge 8, 361 |
| 356, 358, 360, 382 | unsaturation (of lipid acyl chains) 180, 186, |
| aminoacylation 353, 355, 360 | 188, 191–192 |
| anticodon 147-148, 150-151, 352, 355, | Unwin N 8 |
| 358–360, 368, 371–373 | uracil 87, 112–113, 121–122, 126, 129–130, 296 |
| anticodon stem and loop (ASL) 352, | 300, 307, 359, 548 |
| 359, 371–372 | urease 65 |
| CCA end 358, 371, 376 | Urey H 6 |
| cloverleaf 148, 352 | UTR (untranslated region) see mRNA |
| D-stem and loop 148-151, 352, 372 | U-turn see RNA (structural motifs) |
| initiator (fMet) tRNA 362, 368 | UvrB 276 |
| L-shape148–151 | |
| mimicry 378, 381–382 | vacuoles 3 |
| modified bases 128, 130, 148 | validation 24, 555 |
| synthetases, see aminoacyl-tRNA | van der Waals' interactions 38, 168 |
| synthetases | variola (smallpox) 536 |
| tRNA-mRNA interaction 351–352, | VAST 560 |
| 361–362, 367–373, 377, 381 | VCAM-1 512 |
| cognate 356 | vesicle vesicle-mediated transport 194 |
| non-cognate 356 | vesicular stomatitis virus 546 |
| wobble base pairing 125–126, 352, | viruses |
| 373 | assembly 538-542, 547-548 |
| T-stem and loop 149-151, 352 | coat proteins 47–49, 54, 536–537, |
| V (variable) loop 148–151, 352 | 540–541, 546 |
| tropocollagen 507–508 | enveloped viruses 535, 537 542, |
| tropomyosin 55, 58, 491, 494–495, 499, | 544 |
| 501 | entry mechanisms 546–547 |
| muscle 494 | virus-cell fusion 194 |
| non-muscle 494 | visual system 452, 472–477 |
| troponin C, I, T 494–495, 497 | rod cells 472–473, 475 |
| trypsin <i>see</i> proteases (serine proteases) | vitamins 161, 436–437 |
| tubulin 54, 502–505 | B12 |
| tumor necrosis factor 48–49 | K 161 |
| turnip yellow mosaic virus (TYMV) 142 | vitronectin 512 |
| type II diabetes 59 | voltage-gated channels see ion channels |
| v 4 | |

tyrosine kinase see protein kinases

Walker A motif 242
Walker B motif 243
WASP (Wiskott-Aldridge syndrome protein)
485–486, 490–491
water molecules 20, 37, 39, 41–42, 53, 61–62,
70, 123, 135, 143, 148, 151–152, 181, 231–232,
240, 253, 259, 264, 344, 362, 371, 373–374,
417, 429–431, 435, 448, 467, 472, 495, 514–518

bound 41 internal 41–42

Watson JD 7, 105–109 Watson-Crick base pairing *see* base pair WH2 (WASP homology 2) domain 485–486, 488–489, 491

white blood cells 522, 524, 527 Wilkins M 7, 107–108 Willstätter R 8 winged helix-turn-helix motif 321 Woese C 152, 351

xenon 39, 66 X-ray crystallography 109, 140, 444, 555, 568 X-ray diffraction 69, 106, 202 X-ray scattering 169, 189, 208

Z-disc see muscle Z-DNA 109, 116 zeolite-like 169 zidovudine Zif268 see transcription factors (gene-specific) zinc finger 321–323, 325, 394, 486 Z-score 560–561

Index

| AAA+ proteins 245, 251–252, 254, 275, | acyl carrier protein see fatty acid synthase |
|---|--|
| 278–280, 286, 392, 405–416 | acyltransferase 195 |
| ABC transporters <i>see</i> transporters | adaptive immune system <i>see</i> immune system |
| Aβ protein 59 | adaptive initialite system see initialite system adaptor protein 63, 315, 407, 458 |
| Abri 59 | adenine 105, 107–115, 124, 143–147, 156–157, |
| accessible surface area 51 | 295, 300, 326, 347, 361, 406, 548 |
| | |
| actin 48, 53–54, 62, 252–253, 393, 481–502, 504, 511, 561 | adenosylcobalamin 233, 235 adenovirus 540–541 |
| | |
| barbed end 482, 484, 488, 490–491, 494 | adenylate cyclase 468, 476–478 |
| branching and crosslinking proteins see | adenylate kinase 49 |
| Arp2/3 complex | adenylyl cyclase <i>see</i> adenylate cyclase aerobic metabolism 7 |
| capping proteins 482 | |
| fiber formation <i>see</i> formin and spire | aerobic respiration 295 |
| F (filamentous) actin 54, 481–483, 495 | aflatoxin B ₁ 164–165 |
| G (globular) actin 48, 54, 481, 483 | aggregation 32–33, 48, 52–54, 59–62, 168, 182, |
| pointed end 482, 488, 490, 496 | 269, 388–392, 401–402, 533 |
| severing proteins see gelsolin | Agre P 429 |
| active site 48, 50–51, 140, 152, 154–156, | aldolase 49 |
| 227–233, 236–240, 248–254, 260–265, 277, | alphavirus 544–546 |
| 283, 289, 292–297, 300, 327–328, 331–334, | all-α proteins 42–51, 73, 79–95, 559 |
| 348, 340, 342–347, 354, 358, 389–390, | all-β proteins 42–51, 73, 96–102, 427–428, |
| 404–405, 409–419, 435, 457, 460–463, 467, | 559 |
| 472, 479, 482, 570, 574 | allostery 47, 62, 229, 233–239, 393, 401, 463, |
| acetylation see protein modifications | 559 |
| acetyltransferase | α/β proteins 46, 49, 234, 405, 559 |
| in fatty acid synthase see fatty acid | α + β proteins 42–51, 559 |
| synthase | α-crystallin domain (ACD) 391–392 |
| histone modifications see histone | α-helix <i>see</i> helix |
| modifications | α-hemolysin 428 |
| acid base catalysis 153, 229 | Alzheimer's disease 59, 206, 391 |
| active transport 86, 90, 425-426, 443-446 | amino acids 14-19 |
| | |

amino acid sequences 37-40, 43, 50, 78, 81, ATP synthase (ATPase) 241, 243, 245–256, 219, 269, 321, 385, 401, 419, 468, 482, 562, 438 aminoacyl-tRNA synthetases (ligases) (aaRS) 353-361 F₁ 248-255 F_o 248–255 aa-AMP 353, 358 classes 354-356 structure 248-251 subclasses 354-356 editing 359–361 Avery O 7, 107 tRNA recognition elements 356–359 A-minor motif see RNA (structural motifs) amphiphiles 168, 177-185 amphipathic character 70–71, 99, 163 amphipathic helices see membrane proteins amyloid 33, 59-62, 386 amyloid- β (A β) 59 amyloid diseases 59 amyloid-forming proteins 59-63 535–536, 548 Anfinsen C 38, 385, 388 antibodies see IgG anticodon see tRNA antigen-binding domain(s) 522 HK97 540-541 antigens 521–525, 527–528, 533 antigenic peptides 405, 416, 527, 530–531 antiporters see transporters 559 antitrypsin see protease inhibitors apolipoproteins 206, 208–209 BAR domains 199-200 apoptosis 223, 323, 418, 453 aptamer 142–143 base pair aquaporin 52–53, 87, 91, 429–430 selectivity filter 426, 429–432 classification 125 archaebacteria 166, 244, 361 Arf see G-proteins arginine finger 245, 253–254, 277, 465, 467 mismatches 123 Arp2/3 (actin related proteins 2 and 3) 458, 485, 490–493 ASF1 273-734 aspartate carbamoyltransferase (aspartate transcarbamylase) aspartyl proteases see proteases ataxins 58 372-373 atherosclerosis 206 atomic force microscopy 202 ATPases 81, 241, 245, 251, 253 ATP-binding domain 354, 356, 393

binding change mechanism 252 catalytic mechanism 252-254 ATP synthesis 246–247, 252–255 bacteria 3, 5, 69, 79, 90, 96–97, 101–102, 113, 164, 166, 192, 213–214, 218, 220, 222, 224, 245, 247, 249, 251, 256, 260, 275, 279, 283–284, 286–287, 298, 301, 308, 320, 325, 329, 342, 351–352, 361–362, 364, 366–369, 379–381, 388–389, 392, 396, 402–403, 407, 410, 417, 427–428, 444, 446, 481, 521, bacterial chemotaxis 313 bacteriophage 39, 49, 109, 287-288, 292, 318–319, 325, 541, 559 MS2 134, 540–541, 547–548 T4 39, 235–236, 287–288, 537, 548–550, bacteriorhodopsin 8, 71–72, 427, 433, 438–439, base excision repair (BER) see DNA repair conformational parameters 109–114 Hoogsteen (H) 124–128, 136, 151, 295–296 non-Watson-Crick 123-126, 135-136, 304 reverse Hoogsteen 124-126, 136, 149, reverse Watson-Crick 125-126 reverse wobble 126 Watson-Crick (WC) 114-115, 122-126, 132, 134, 141, 150–151, 294, 297, 304, 307, wobble base pair see also tRNA-mRNA interaction 125–126, 373 base quadruplets 127–128

| bases (in nucleic acids) see nucleosides | C1q see complement system |
|--|---|
| base triplets 127–128 | Ca ²⁺ -ATPases <i>see</i> transporters |
| codons of mRNA 352-353 | cadherin see CAM |
| RecA 303-304 | calcium-gated potassium channel see ion |
| basic-helix-loop-helix (bHLH) proteins | channels |
| 317–318 | calmodulin 497 |
| Bawden FC 8 | calorimetry 169, 189, 202 |
| Bence-Jones proteins 523 | chameleon sequences 565 |
| Bernal JD 8 | cAMP 427, 433, 468, 478 |
| Berzelius JJ 7 | capping proteins see actin |
| β-adrenergic receptor see G-protein coupled | carbohydrates 3, 7–9, 213–224 |
| receptors | monosaccharides 213–216, 219, 221 |
| βαβ units 46 | deoxyhexose 215 |
| β-bulge 33–34, 101 | hexosamine 215, 221 |
| β-2 microglobulin 59, 528–529 | hexose 215 |
| β hairpin 43, 321, 455 | galactose 215 |
| β-helix or solenoid 43 | glucose 215–218, 220–222, 322, |
| β-sheets | 425–426, 444, 483 |
| antiparallel 32–33, 43–44, 325, 337, 355, | mannose 215, 220 |
| 377–378, 458–459, 479, 491, 497, 522, | pentose 109, 215 |
| 547 | ribose 109–115, 121–124 |
| barrel or cylinder 43–44, 46, 49 | xylose 215 |
| Greek key 43–44, 97 | uronic acid 215, 221, 223 |
| Jellyroll 43–44, 48–49, 540–541, 546 | sialic acid 215 |
| meander 43–44, 97, 491 | oligosaccharides 213, 216, 219-220, 223, |
| parallel 43–44, 46, 49, 242, 257, 369, 462, | 427 |
| 483, 504 | non-reducing end 216 |
| propeller 43–44, 338, 513 | reducing end 216 |
| sandwich 43–44, 48–49, 391, 454–455, 458, | polysaccharides 163, 213, 221–224 |
| 510, 522–523 | carbonic anhydrase 49, 51, 63, 229–233 |
| twist 32–33, 59, 98 | active site 231 |
| up-and-down 43–44, 97–98 | sulfonamide inhibitors 231 |
| β-turn 33–34 | cardiolipin see glycerophospholipids |
| B-factor 419, 556–557 | carotenoids 162 |
| binding change mechanism see ATP synthase | Caspar D 8, 537–538 |
| biotin 142–143, 256 | catalytic mechanism 50-51, 153-155, 222, 229 |
| BLAST 564 | 233–234, 241, 252, 254, 290, 294, 342, 388 |
| blood coagulation 224, 404, 418, 455 | catecholamines |
| BLOSUM62 570 | CATH see fold databases |
| bluetongue virus | caveolae 205 |
| bond energies 12 | caveolin 205 |
| bond lengths 12 | CD155 see poliovirus receptor |
| Boyer PD 252 | CD3 528, 533 |
| bracket notation 141–142 | CD4 528, 531–532, 543 |
| bulge see β bulge or RNA (structural motifs) | CD8 528, 531–322 |

chemiosmotic theory 245, 247

chirality 15, 29

consensus sequence (sequon) 134, 141,

219–220, 308, 324, 330, 337, 354, 465

| cooperativity 399–400, 417 | succinate dehydrogenase 246 |
|--|--|
| negative 400 | Deisenhofer J 72 |
| positive 400 | denaturation see protein denaturation or |
| CorA see ion channels | nucleic acid denaturation |
| cotransporters 425 | dengue virus 545 |
| corneocytes 175 | dephosphorylation 435, 452, 461 |
| CORN rule 15, 25 | depolymerization 54, 490 |
| Coulomb's law 13 | detergent 69, 72, 76, 166, 198, 201, 208 |
| covalent bond 11-13, 20, 228-229, 273 | dextro 15 |
| Creutzfeldt-Jakob's disease 59 | diacylglycerol see glycerolipids |
| Crick FHC 105–109, 125, 148 | dielectric constant 13 |
| cro repressor see transcription (bacterial | differentiation 224, 307, 526-527 |
| repressors) | diffusion rate 426 |
| cross-β 59–61 | disorder |
| cross-link 40, 63, 136, 482, 490, 493–494 | in lipid membranes 171, 173, 181, |
| Crowfoot-Hodgkin D 8 | 202–204 |
| cryo-EM 8, 271–272, 280–281, 366, 400, 415, | in proteins 47–48, 59, 239, 321, 330, |
| 495, 499, 536, 545, 550 | 456–457, 471, 483, 487, 504, 540–544, |
| crystallization 7–9, 52, 69, 72, 151, 159, 181, | 556, 567–569 |
| 469, 483, | distance plot 554 |
| crystallography see X-ray crystallography | disulfide bond 13, 19, 240, 388-389, 455, 462, |
| CTP:phosphocholine cytidyltransferase (CCT) | 509, 523 |
| 197–198 | divergence see evolution |
| cubic phase 174, 179, 181, 191, 195 | DNA (deoxyribonucleic acid) |
| cyclic GMP phosphodiesterase (PDE6) | 30 nm fiber 269, 271 |
| 473–475 | A-DNA 106–107, 116 |
| cyclophilin 386–388 | B-DNA 106, 109, 119, 303–304 |
| cyclosporine A 386 | backbone angles 109–114 |
| cysteine-rich domain 455, 458, 461–462 | double helix 108-109, 114-120, 122, 272, |
| cytochrome bc ₁ complex 246 | 297 |
| cytochrome c oxidase 85, 246 | double-stranded (ds, duplex) 111, 114, 116 |
| cytosine 105–109, 112–113, 115, 121, 153, | 121–122, 148, 269, 273–274, 277, 280–281 |
| 295–297, 300, 313, 316, 326 | 284, 288–289, 291–292, 326–327, 333–334 |
| cytokine receptors 453, 455, 461 | 338, 342–343 |
| cytoskeleton 481–482, 489, 496, 514 | major groove 115-116, 316-325, 337-338, |
| cytosol 3, 13, 71, 74–75, 92–93, 95, 98, 101, 396, | 340 |
| 401, 435, 440, 528 | minor groove 115–116, 291, 294, 297, 321, 324–325, 337–338 |
| Dali 554, 560 | non-coding 137 |
| deacetylation 315 | syn/anti conformation 113–114, 157, 296 |
| dehydratase see fatty acid synthase | Z-DNA 109, 116 |
| dehydrogenases 46–47, 52–53, 246, 503 | DNA-binding domain (DBD) 311-312, |
| lactate dehydrogenase 47, 53 | 316–317, 321–324, 456 |
| malate dehydrogenase 47 | DNA-binding proteins 115, 302, 321 |
| NADH dehydrogenase 246 | DNA damage 301, 346 |

| DNA helicases see helicases | ectodomain 511, 513–515 |
|--|---|
| DnaB family see helicases | effector site 237–238, 240 |
| DnaJ see heat shock proteins | EF-Tu, EF-G etc. (elongation factors) see |
| DnaK see heat shock proteins | translation |
| DNA ligase 273, 275 | EF hand 458, 494–497 |
| DNA polymerase | EGF see epidermal growth factor |
| classes 287–288 | EGF receptor (EGFR) 462 |
| exonuclease domain 273, 289-295 | elastin 224 |
| fingers domain 290–292 | electrochemical (or ion) gradients 69, 245-246, |
| Klenow fragment 288–294 | 255–256, 425–426, 434 |
| palm domain 289–293, 295 | electron diffraction 8, 71 |
| primase 288, 290, 294 | electron microscopy 8, 71, 162, 206, 275, 364, |
| primer (DNA or RNA) 273, 276, 286-287, | 366, 445, 483, 498, 548, 555 |
| 289–294 | electron transfer 64-65, 81, 274 |
| reverse transcriptase (RT) 288, 299, 300, 347 | electrostatic interaction 13, 31, 76, 197, 229, 247 |
| thumb domain 290, 292 | elongation factors see translation |
| DNA repair 273, 276–277, 288, 295–297, 323 | endoplasmic reticulum (ER) 93–94, 194, 197, 220, 388, 390, 481, 528 |
| base excision repair (BER) 297 | endosome 543 |
| 8-oxoG-DNA glycosylase (OGG1) | enoyl reductase see fatty acid synthase |
| 296–297 | enthalpy 14, 40 |
| nucleotide excision repair (NER) 297 | entropy 37, 39–40, 205, 247 |
| DNA-RNA hybrid see RNA polymerase | epidermal growth factor (EGF) 219, 462–465, |
| DNA synthesis see replication | 513–514 |
| DNA template 108, 114, 146, 272–276, 285–286, | module 514 |
| 288–295, 298–301, 307, 326, 328, 332–335, | receptor 462–465 |
| 340, 342–345 | epigenetics 271, 308, 312, 316 |
| DNA topoisomerase 128, 277, 282–285 | epinephrine (adrenalin) 451, 471, 475–477 |
| single strand (Topo IA, Topo IB) | epinephrine receptor 451, 468–469, 475–478 |
| 282–283 | ERK (extracellular-regulated kinase) see |
| double strand (Topo IIA, Topo IIB) | protein kinases (Ser/Thr kinases) |
| 284–285 | essential light chain see myosin |
| docosahexaenoic fatty acid 164 | etioplasts 194 |
| domain swap 54, 61, 395 | eukaryotes 6, 13, 79, 90, 93, 96, 152, 166, 214, |
| double bond 12, 20, 164–166, 186–187, 205, | 218–220, 244, 269, 272–276, 280, 283–288, |
| 387 | 294, 301, 307–308, 316, 325, 327, 329, 331, |
| drugs 69, 170, 175–176, 224, 282, 386, 395, 446, | 351, 361–362, 367–369, 377, 379–380, 388, |
| 475, 503, 546 | 390–392, 401, 403, 407, 413, 417, 430, 436, |
| drug delivery 170 | 488, 521 |
| proteins as drug targets 69, 446, 503 | evolution 1–10, 15, 43, 48–51, 55, 90, 121, |
| DsbA, DsbB, DsbC, DsbD disulfide bond | 233–234, 287, 301, 351–353, 386, 482, 498, |
| formation proteins 388–389 | 528, 559, 572–573 |
| dynamics see protein dynamics | convergent 51, 233 |
| dynein 503 | divergent 234 |

| exonuclease domain see DNA polymerase | proteins 37–41, 52, 59, 93, 101–102, 199, |
|---|---|
| extracellular matrix (ECM) 213–214, 223–224, | 219–220, 385–411, 559, 562, 565, 571 |
| 507–508, 511 | spontaneous 385–403 |
| extracellular signals 452 | assisted 386, 403 |
| eye lens 391, 573 | RNA 125, 137, 147–151, 156, 159 |
| | fork loop see RNA polymerase |
| F ₁ see ATP synthase | formin 485–488 |
| Fab see IgG | FH1 485, 487–488 |
| F-actin see actin | FH2 485, 487–488 |
| familial amyloidotic polyneuropathy 59 | lasso, linker, knob, coiled-coil, post |
| familial British dementia 59 | 487 |
| fatty acids 163–165, 188, 190, 215, 260, 262, | Fos see transcription factors (gene-specific) |
| 264–265 | four-helix bundle 46, 318, 454 |
| in lipids 161, 163–165, 188, 190 | Franklin, R 7, 105–107 |
| fatty acid synthase (FAS) 197, 260–265 | frustration (membrane) 180-182 |
| Type I 260 | FSI (formin spire interaction) 487–488 |
| Type II 260 | FtsH see proteases (oligomeric ATP-dependent |
| ferredoxin 49, 66, 262 | proteases) |
| ferritin 53, 65 | furin-like see cysteine-rich domain |
| FeS clusters 65 | fusion peptide 195–197, 542–546 |
| FfH signal recognition protein | |
| fiber diffraction 59, 105, 483 | G-actin see actin |
| fibrin 224, 508–509 | γ turn 23 |
| fibrinogen 512 | GAP (GTPase-activating proteins) 257–259, |
| fibrinolysis 418 | 373, 458, 465, 467, 472, 484 |
| fibronectin 224, 454–455, 461, 508–512, | RasGAP 458, 467 |
| 516 | GAL4 see transcription factors (gene-specific) |
| modules in proteins 508–510 | gated channels 426–427, 430 |
| type I 509–510 | GCN4 see transcription factors (gene-specific) |
| type II 509–510 | G-domain see G-proteins |
| type III 509–510 | GEF (guanine nucleotide exchange factor) |
| filamentous actin see actin | proteins 257, 374, 378, 458, 465–466, 468, 471, |
| finnish-type familial amyloidosis 59 | 474 |
| FK506-binding proteins 386 | gel phase 171, 173, 175, 186, 203 |
| flavivirus 140, 536, 544–545 | gelsolin 59, 485, 488–490 |
| fluid mosaic membrane model 199–201 | gene |
| F _o see ATP synthase | deletion 301 |
| fold classification 559 | duplication 1, 42, 90, 429, 489, 573 |
| fold databases 559–560 | expression 9, 113, 118, 137, 219, 270, 273, |
| CATH 43-44, 559-560 | 298, 301, 307–308, 323, 391, 452, 465, 491, |
| SCOP 43, 559–560 | 526 |
| fold recognition 569–571 | fusion 42, 575 |
| folding | insertion 564 |
| funnel 38 | genetic code 4, 120, 126, 147–148, 246, |
| process 37–39, 49, 385–386, 403 | 351–352 |
| | |

| codons 125, 129-130, 146-147, 150, 352, | glycosyl hydrolases 216–218 |
|--|--|
| 362, 367–368, 372–374, 377, 379–380, | glycosyl transferases (GT) 216–219 |
| 547–548 | Gly-zipper 55 |
| degeneracy 120, 126, 147-148, 352, 391, 411 | G-nucleotide exchange factor see GEF |
| genetic exchange (crossing over) 117, 301 | Golgi complex 3, 194, 197, 219–220, 528 |
| GFP (green fluorescent protein) 97, 491 | GPCR see G protein-coupled receptors |
| Gibbs energy 179–180, 247 | GPI (glycosylphosphatidylinositol) anchor 73 |
| Gibbs phase rule 170 | G protein-coupled receptors (GPCR) 427, |
| globin fold 47, 49 | 467–471, 475–478 |
| glucagon 451 | adrenergic receptor 468–469, 475–478 |
| glucagon receptor see G protein-coupled | glucagon receptor 468 |
| receptors (GPCR) | glutamate receptors 468–469 |
| glucocorticoid receptor see transcription | olfactory receptors 468 |
| factors (gene-specific) | pathways 468, 470, 476, 478 |
| glucokinase 561 | rhodopsin 427, 468–475, 477 |
| glucomannokinase 561 | G-proteins 205, 241, 243, 257–259, 368–370, |
| glucose see carbohydrates | 377, 452, 465, 468, 470–472, 477, 484, 491, |
| glucose transporters 425 | 498, 500, 503, 512, 514 |
| glutamate receptors see G protein-coupled | GAP see GAP |
| receptors | G-domains 257–259, 471 |
| glutaredoxin 235 | GEF see GEF |
| glycerol backbone 163–164, 166 | GTP hydrolysis 257–260, 371–374, 465, 467 |
| glycerol kinase 561 | Ras superfamily of monomeric G-proteins |
| glycerolipids 163–165 | 452 |
| diacylglycerol 195, 197 | Arf 452 |
| diglucosyldiacylglycerol (DGlcDAG) | Rab 452 |
| 189–190 | Ran 452 |
| monoglucosyldiacylglycerol (MGlcDAG) | Ras 49, 243, 257–258, 368, 370–371, |
| 174, 189–190, 192 | 452, 464–467, 471–472, 498, 554 |
| glycerophospholipids (phospholipids) | Rho 452 |
| 163–166, 168, 173, 202–203, 207–209 | trGTPases 363, 368–369, 373, 377 |
| cardiolipin 166 | trimeric G-proteins 243, 257, 370, 452, 458, |
| diphosphatidylglycerol 166, 188 | 468, 470–473, 476–478, 491, 512, 514 |
| phosphatidylcholine (PC, lecithin) 70, 75, | Gα 471, 473, 514 |
| 166, 168, 171, 179–180, 194, 197 | Gβ γ 470–471 |
| phosphatidylethanolamine (PE) 13, 75, | transducin |
| 188, 195–197 | grana stacks 194 |
| phosphatidylglycerol (PG) 75, 166, 188, 190 | Grb2 adaptor protein 458, 465 |
| phosphatidylinositol 166 | Greek key motifs see β -sheet |
| glycolipids 75, 167, 214, 218, 530 | green fluorescent protein see GFP |
| glycoproteins 213–214, 218–221, 224, 437, 526, | GroEL/GroES (GroE) 52–53, 391, 396–403 |
| 530, 536 | bullet model 398–400 |
| glycosylation 90, 218–220 | football model 399–400 |
| O-glycosylation 219 | refolding of misfolded proteins 391–392, |
| N-glycosylation 220 | 405 |

| growth hormone (GH) 451, 453–455 | helical net diagrams 57–58 |
|---|--|
| growth hormone receptor 453–454 | helical wheel diagrams 56–57, 78 |
| GrpE 391, 393, 561 | helicases 243, 245, 273, 275–287, 335 |
| G-tetrads see RNA (structural motifs) | DNA helicases 243, 276–282 |
| GTP hydrolysis <i>see</i> G-proteins | hexameric 243, 275, 278, 280–281 |
| GTPases see G-proteins | SF3, SF4, SF5, SF6 276–282 |
| guanine 105, 107–109, 112–115, 124, 126–129, | monomeric |
| 146, 154, 242, 295–296, 300, 324–325, 347 | SF1, SF2 277 |
| GxxxxGKT/S sequence motif (P-loop) | RNA helicases 276 |
| 242–243, 252–254, 257–259, 276–277, 301, | DEAD-box helicases 276 |
| 369–370, 374–379, 408, 467, 472, 498, 503 | helicase loader 243, 273, 275–276, 279–280, 286 |
| GxxxG sequence motif 83-84, 462 | helix-helix interactions 81–84 |
| • | helix-loop-helix motif 291, 317–319 |
| H ⁺ -ATPase <i>see</i> transporters | helix-turn-helix motif 284, 316–321, 330 |
| H, K ⁺ -ATPase <i>see</i> transporters | hemagglutinin see influenza virus |
| Haldane JBS 6 | heme 40, 63, 65 |
| halophilic 166 | hemerythrin 49, 53 |
| hammerhead ribozyme 138–139, 153–157 | hemicellulose 214, 222, 224 |
| handedness 29–30, 46, 56, 81, 97–98 | hemoglobin 7, 8, 42, 49 |
| left 24–25, 29–31, 56, 81, 83, 98, 109, 116, | Henderson R 8, 71 |
| 248, 269, 281, 303, 309, 459, 530 | heparan sulfate 221, 223–224, 508 |
| right 24, 29–31, 46, 55–56, 81–82, 97–98, | heparin 221, 223–224, 509–510 |
| 109, 115, 243, 279–281, 286, 303 | heptad repeat 55–57, 78, 317, 494 |
| haptens 524 | hereditary systemic amyloidosis 59 |
| heat shock proteins <i>see also</i> chaperones | herpes simplex virus 546 |
| HslVU see protease | hexagonal phase 171, 174, 180 |
| Hsp10 see GroEL/GroES | normal (H _I) 171, 174, 177 |
| Hsp40 (DnaJ) 391–395, 397, 402–403 | reversed (H _{II}) 171, 174, 177, 179–182, 188, |
| Hsp60 see GroEL/GroES | 193, 195, 197 |
| Hsp70 (DnaK) 391–397, 402–403, 408, 482, | hexokinase 482–483, 561 |
| 561 | Hidden Markov Model 571, 573 |
| Hsp90 391–392, 395–397 | histone acetyltransferases (HAT) 314 |
| Hsp100 (ClpA) see proteases | histone deacetylases (HDACs) 314 |
| Hsp104 (ClpB) see proteases | histone demethylases (KDM) 313 |
| Hsp110 391–392 | histone methyltransferases (HMT) 313 |
| small heat shock proteins (sHsps) | histones |
| 390–393 | H1 271, 316, 413 |
| helix | H2A 269, 273 |
| α-helix 24, 27–30, 46, 55, 57, 78, 80, 84, 206, | H2A.Z 310, 312 |
| 292, 316, 322, 453, 471, 517, 558, 563 | H2B 269, 273 |
| 3 ₁₀ helix 30 | H3 269–270, 273–274, 315, 338–339, 413 |
| helical bundle 65, 81, 90, 345, 407 | H4 269–270, 273–274, 314, 338–339, 413 |
| helical dipoles 30–31, 86–87, 441 | H5 270–271, 316 |
| π helix 30, 84 | histone code 271, 316 |
| poly-Pro helices 31, 459–460, 530 | histone fold 269-270, 338-339 |
| | |

| histone modifications 271, 273, 311–315 | 357–359, 370–374, 376, 388, 410, 420, |
|---|---|
| arginine methylation 313, 315 | 429–430, 472, 508, 516, 567 |
| lysine acetylation 313–315 | hydrogenase (Ni-Fe hydrogenase) 64 |
| erasers312, 315 | hydrophilic character 41, 65, 161, 168, 182, 220, |
| lysine metylation 313 | 398, 410, 420 |
| readers 312 | hydrophobic character |
| Bromo 315 | core 13, 37, 39, 42, 50, 56–57, 73, 75, 86–87, |
| PHD 315 | 98, 143, 566 |
| writers 312, 315 | effect 37-41, 52, 168, 182 |
| N-terminal tails 269–270 | (non-polar) interactions 14, 18, 37, 76–78, |
| octamer 269–270, 309, 334, 338 | 80–81, 83, 161 |
| HIV (human immunodeficiency virus) 132, | surface 389, 407, 430, 463 |
| 135, 137, 213, 387–388, 524–526, 536, 543, | hydroxyproline 31, 508, 512 |
| 559 | hypertension 206 |
| Env (=gp120 + gp41)525, 543 | Hünefeld FL 7 |
| gp120 525–526, 532, 543 | |
| HIV reverse transcriptase see DNA | I domain see integrins |
| polymerase (reverse transcriptase) | ICAMs (intracellular adhesion molecules) 512 |
| Rev protein 135 | icosahedral symmetry see symmetry |
| HLA see MHC | icosahedral viruses 54, 538, 540 |
| HNF-3 see transcription factors (gene-specific) | IgA (immunoglobulin type A) 522 |
| Hodgkin DC see Crowfoot Hodgkin | IgD (immunoglobulin type D) 522 |
| Holliday junction 117, 561 | IgE (immunoglobulin type E) 522 |
| homeobox sequences 320 | Ig fold (immunoglobulin fold) 48, 454, 458, |
| homeodomains see transcription factors | 511, 522, 533 |
| (gene-specific) | IgG (immunoglobulin type G) 522–523 |
| homologous proteins 49–50, 569, 572–573 | broadly neutralizing antibodies (bnAbs) |
| homologous recombination (HR) 297, 301, | 524–526 |
| 305 | CDRs (complementarity-determining |
| presynaptic filament 302–304 | regions) 522–524, 527, 530 |
| homology modeling 324, 569–571 | constant domain 522–523 |
| hormones 161–162, 164, 166, 322, 453 | Fab 522–524, 530 |
| HSlU see protease | Fc 522–523 |
| HSIV see protease | neutralizing antibodies (nAbs) 524–524 |
| Hsp proteins see heat shock proteins | variable domain 522–524, 530–531 |
| Huber R 73 | IgM (immunoglobulin type M) 522, 527 |
| humoral immunity 521 | immune system 48, 428, 451, 455, 521–533 |
| huntingtin 59 | innate 428, 521 |
| Huntington's desease 59 | non-adaptive 521 |
| hyaluronan 213–214, 221, 223–224 | antibody-mediated 521–527 |
| hydrogen bond 12–15, 18, 27–34, 37–41, 52, | T-cell mediated 527–533 |
| 57, 61–62, 74, 81, 83, 85–87, 97, 100, 107–108, | immuno-proteasome 416 |
| 114–115, 122–127, 133–136, 145, 150–155, | inclusion bodies 386, 391 |
| 157, 159, 168, 222, 231–232, 239–240, 257, | inflammation 418 |
| 270, 292, 294, 304, 312, 314–315, 324–325, | influenza virus 52, 213, 524, 536, 543–545 |

| hemagglutinin 52, 196, 536, 543, 545 neuraminidase 52, 524, 536 | kinesins C-, M- (or KinI) and N-type 498, 503–505 |
|--|--|
| initiation factors see translation | ncd (non-claret disjunctional) 505 |
| initiator tRNA see tRNA | kissing loops 132–133 |
| in-line S _N 2 attack 259–260 | Klenow fragment see DNA polymerase |
| inosine 112, 130, 148 | Klug A 8, 537–538 |
| inositol 166 | Kornberg A 288 |
| insulin 39, 572 | Kornberg R 330 |
| | ~ |
| insulin receptor 458, 461–462, 464 | Kv1.2 <i>see</i> ion channels (potassium channels) |
| integral membrane proteins see membrane | Citatitieis) |
| proteins | la stata dahardua annaga saa dahardua annaga |
| integrase 561 | lactate dehydrogenase <i>see</i> dehydrogenases |
| integrins see CAM | lamellar phase 174–176, 194, 202–204 |
| intercalation 337 | laminins 512 |
| interfacial region 70–71, 76, 87–89, 99, 168, | lateral pressure 182–185 |
| 566–567 | lectins 213 |
| interfacial tension 182, 184 | leucine-rich repeats 462 |
| interferons 451, 453 | leucine zipper 196, 317–319 |
| introns see mRNA | LeuT transporter 90, 440–441, 446–449 |
| ion channels | lever rule 173 |
| cGMP-gated sodium channels 473–474 | Levinthal C 38, 385 |
| Mg ²⁺ channel CorA | levo 15 |
| potassium channels 430–433 | light-chain amyloidosis 59 |
| KcsA 430–433 | light-harvesting complex 52 |
| voltage gated K ⁺ channels (Kv1.2) | lipids |
| 431–433 | aggregate structures 161–162, 168, |
| MthK 433 | 177–187, 195 |
| selectivity filter 426, 429–432 | categories 163–164 |
| ionotropic receptors see transporters | lateral diffusion 185–187, 189, 203–205 |
| islet amyloid polypeptide (IAPP) 59–60 | rafts 173, 186, 199–206 |
| isosteric 126, 149 | synthase 197 |
| | vesicles 77, 176–177, 179, 193–197, 199, |
| JAK Janus kinase family see protein kinases | 202, 481, 496, 502–503, 528 |
| (tyrosine kinases) | large unilamellar (LUV) 176–177 |
| jellyroll fold see β-sheet | small unilamellar (SUV) 176 |
| Jun see transcription factors (gene-specific) | lipoproteins 3, 206–208 |
| | LDL 206–208 |
| K ⁺ channels see ion channels (potassium | HDL 208 |
| channels) | liposome 162, 176 |
| KcsA see ion channels (potassium channels) | lyotropic liquid crystals 169 |
| keratin 42 | lyotropic polymorphism 169–170 |
| ketoacyl reductase see fatty acid synthase | liquid crystalline phases 168–170, 173–174, |
| ketoacyl synthase see fatty acid synthase | 178–179, 188 |
| KIM (kinase interaction motif) 465 | liquid-disordered (l _d) 204–205 |
| kinase domains 455, 459, 461–463 | liquid-ordered (l _o) 204–205 |

α-helical 79–84, 90

| LonA/B see proteases | multiple pass or polytopic |
|--|--|
| LUV see vesicles | 74–75, 79, 81, 84, 90–91, |
| lysidine 129–130 | 94–95 |
| lysine oxidase 508 | inverted repeat 90–91, |
| lysosomes 3, 404 | 429, 440–441 |
| lysozyme 39, 42, 59, 222, 524, 549–550, 559, | non-inverted repeat 91 |
| 573 | single pass <i>or</i> bitopic 73–75, |
| | 79, 81, 83, 90–95 |
| MacLeod C 7 | β-barrel <i>or</i> outer membrane |
| mad cow disease 61 | proteins (OMPs) 96–102 |
| major groove see DNA or RNA | peripheral membrane proteins 72–74, |
| major histocompatibility complex | 75–78, 89, 102, 194, 199 |
| see MHC | amphipathic helices 76–78, |
| malonyl/palmitoyl transferase see fatty acid | 197–198, 491 |
| synthase | conditional 73, 76 |
| Mat a1 see transcription factors (gene-specific) | monotopic 73 |
| Mat α2 see transcription factors (gene-specific) | Meselson M 272 |
| MAP kinase see protein kinases (Ser/Thr | messenger RNA see mRNA |
| kinases) | metabolism 1–2, 5, 7, 205, 219, 272, 315, |
| maturation cleavage 546 | 322 |
| Max see transcription factors (gene-specific) | metabotropic glutamate receptor family see |
| M-band see muscle | G protein-coupled receptors |
| McCarty M 7 | metabotropic receptors 427 |
| mechanosensitive channels 185, 433 | metal binding 63–66, 514–515 |
| Mediator see transcription factors (general) | metal clusters 65 |
| medin 59 | metal ion dependent adhesion site (MIDAS) |
| meiosis 301 | see integrins |
| membrane(s) | metastable proteins 418 |
| biogenesis 197 | methylated bases see nucleosides |
| curvature 77, 179, 195, 198–200 | methylation see protein modifications |
| mean 204 | methyl transferases 313 |
| monolayer 177, 179–181, 185–186 | Mg ²⁺ channel <i>see</i> ion channels |
| spontaneous 177, 179–180, 194–195, | MHC (major histocompatibility complex) or |
| 198 | HLA (human leucocyte antigen) 405, 416, |
| total 179 | 521, 527–533 |
| domain formation 176, 188, 201–202, | class I 416, 527–532 |
| 204–205 | class II 527–532 |
| fission 194–196 | Michel H 72 |
| fusion 176–177, 194–197, 525, 542–546 | Miescher F 7 |
| stalk formation 195 | microfilaments 481, 491, 502 |
| membrane attack complex 428 | microtubules 54, 502–505 |
| membrane proteins | MIDAS motif see integrins |
| integral or transmembrane proteins (TMP) | Miller S 6 |
| 57 73 70_06 566 | minor grooms sag DNA or PNA |

miRNA (microRNA) 307

| misfolded proteins 400, 405 | mutations 19, 39–40, 109, 120, 122, 130, 132, |
|---|---|
| Mitchell P 246–247 | 149, 157, 273, 294–295, 323, 346, 379, 390, |
| mitochondria 13, 96, 98, 101, 161, 166, 245, | 417, 441, 467, 527, 573 |
| 247, 251, 297, 352, 361–362, 396, 410, 427, | Myc see transcription factors (gene-specific) |
| 502 | MyoD see transcription factors (gene-specific) |
| mitochondrial membrane 194, 248 | myoglobin 8, 30, 40, 42 |
| molecular dynamics 324, 431, 441, 571 | myosin I, II, V, VI 496–502 |
| molecular evolution 48–51, 55, 233–234, 287, | actin binding 493–495, 501, 504 |
| 301, 351–353, 559, 572–573 | conformational states 500–502 |
| molecular genetics 301 | rigor, post-rigor, pre-power stroke, |
| molecular motors (see also myosin or kinesin) | power stroke 500–502 |
| 229, 276 | converter domain 497, 500 |
| molecular switches see G-proteins | essential light chain (ELC) 497, 501 |
| molten globule 39, 419 | heavy chain 496–497, 501 |
| monoamine oxidase 79 | regulatory light chain (RLC) 497, 501 |
| monolayer curvature see membrane(s) | S1 fragment 497, 499 |
| motility 481–505 | similarities with G-proteins 498 |
| motor domains, motor proteins see myosin or | myristoylation see protein modifications (lipid |
| kinesin | modifications) |
| mRNA 105, 122, 128, 141, 146–147, 152, 156, | |
| 316, 330, 344, 346–347, 351–352, 361–363, | Na ⁺ ,K ⁺ -ATPase <i>see</i> transporters |
| 366, 368–369, 372–373, 377, 379–381, 502, | Na ⁺ /H ⁺ antiporter 440 |
| 526 | nascent polypeptide see translation |
| coding region 121, 137, 146, 308 | natively unfolded/unstructured protein 48 |
| codon see genetic code | ncd see kinesins |
| codon-anticodon interaction see | neural network 563–565, 568 |
| translation | neuraminidase see influenza virus |
| initiation codon (AUG) 352, 367, 547-548 | neurotransmitters 446, 474–475 |
| introns 139, 152, 156–157, 309, 330, | neutron diffraction 41 |
| 346–348 | NhaA see transporters |
| group I intron 152, 156–157 | Nicolson GL 199, 201 |
| group II intron 152 | nitrogen fixation 64 |
| poly(A) tail 146 | NMR spectroscopy 9, 41, 189, 208–209, |
| stop codon 146-147, 352, 362, 379-380 | 555 |
| UTR (untranslated region) 140, 146 | non-Watson-Crick base pairing see base pair |
| MthK see ion channels (potassium channels) | non-coding DNA see DNA |
| multidomain proteins 257, 260, 458 | non-histone chromosomal proteins 269 |
| multifunctional enzymes 260 | nonlamellar phases 173-174, 188, 190-191, |
| muscle 54, 205, 318, 434, 481–482, 493–501 | 194 |
| contraction 493–494, 498–502 | Northrop JH 8 |
| M-band 493–494 | NSS see transporters |
| sarcomere 493–494, 501 | Ntn-hydrolases 416–417 |
| thick filaments 493-494, 496, 501 | nuclear receptors 317, 322 |
| thin filaments 54, 493-494, 501 | nuclease 38, 271, 385 |
| Z-disc 493–494, 501 | nucleic acid denaturation 118, 276 |

| nucleic acid structure see DNA or RNA | p23/Sba1 co-chaperone <i>see</i> chaperones |
|---|--|
| nucleocapsid 542 | P4-P6 domain see ribozymes |
| nucleophilic attack 154, 156, 231–232, 347, | p53 see transcription factors (gene-specific) |
| 435 | packing parameter 177–179, 195 |
| nucleosides 109–110 | palindromic sequences 319, 321, 325 |
| bases 112-113 | palm domain see DNA polymerase |
| methylated bases 113 | palmitoyllinoleoyl -PC 166 |
| syn/anti conformation 113–114, 157, 296 | palmitoylarachidonyl -PC 166 |
| nucleoside analogs | palmitoyldocosahexaenoyl-PC 166 |
| AZT | Parkinson's disease 59, 391, 446 |
| nucleosome 269–271, 309–313, 316, 338 | partial charges 19, 31 |
| histones see histones | passive transport 248, 425–427 |
| linker DNA 269, 271, 309 | pathogen 527, 530 |
| non-histone proteins 313 | PDB see Protein Data Bank |
| nucleosome-depleted regions (NDR) 310 | PDI see protein disulfide isomerase |
| nucleosome-free regions (NFR) 310 | pepsin see proteases |
| tetranucleosomes 271–272 | peptide bond 20–23 |
| nucleotide exchange factor (see also GEF | cis 20, 22, 31 |
| proteins) 257, 369, 374, 378, 391, 465, 488, | trans 20, 22, 27, 31 |
| 561 | peptidoglycan (murein) 213–214, 224 |
| nucleotide excision repair (NER) see DNA | peptidyl-prolyl-cis/trans isomerase (PPIase) |
| repair | 402 |
| nucleotides 109–110 | peptidyl transfer <i>see</i> translation |
| nucleus 3–4, 297, 308, 323, 330, 451–453, 456, | peripheral membrane proteins <i>see</i> membrane |
| 464–465, 488, 542 | proteins |
| 101 100/ 100/ 012 | Pfam database 573–574 |
| OB-fold 415 | phagocytosis 521–522 |
| Oct-1 see transcription factors (gene-specific) | phase diagram 169–175, 181, 202–204 |
| odorant receptors see G protein-coupled | ternary 172–173, 203–205 |
| receptors (olfactory receptors) | three-phase line 173 |
| odorants 451 | phase transition 169–170, 173, 175–176, 182, |
| Okazaki fragments see replication | 189, 191–193, 203 |
| olfactory receptors see G protein-coupled | PHD 315, 564 |
| receptors | PHDsec 564–565 |
| ω (omega) torsion angle 22 | phosphatases 452, 457–458, 465 |
| OmpF 72, 97, 99 | Phosphatidylcholine 70, 166, 168, 179, |
| Oparin A 6 | 194 |
| O-phosphatidyltrimethylarsonium lactic acid | phosphatidylinositol <i>see</i> glycerophospholipids |
| 168 | (phospholipids) |
| opsin see G protein-coupled receptors | phospholipases 166–168, 194–195, 458, 468 |
| (rhodopsin) | phospholipase C 167, 194, 458, 468 |
| ovalbumin 418 | phospholipids see glycerophospholipids |
| Overton EC 161 | phosphonolipid 167 |
| | |
| oxidative phosphorylation 246–247 | phosphopanteine transferase <i>see</i> fatty acid |
| oxoG glycosylase 296–297 | synthase |

| phosphorylation 215, 246–248, 312, 391, 435, | profile methods 569 |
|--|--|
| 451–452, 457, 459–460, 463, 527 | profilin 484–485, 487–489 |
| photosynthesis 7, 64, 386 | ProFunc 574 |
| photosynthetic reaction center 72, 88 | promoter 128, 308–310, 323, 325–326, 329, 330, |
| phytanyl 166 | 334–342 |
| π-electrons 19–20 | propeller protein 43, 119, 256, 338, 458, 461, |
| Pirie NW 8 | 470, 491, 512–517 |
| pleckstrin (PH) homology domain 458–459 | ProSite 573–574 |
| P-loop motif see GXXXXGKT/S | prostaglandins 164 |
| pol I, pol II, pol III see RNA polymerase | prosthetic groups 63–66 |
| (eukaryotic) | proteases (proteolytic enzymes) 42, 50–51, |
| poliovirus 536, 546–547 | 223–224, 229, 329, 385, 404–407, 409–411, |
| poliovirus receptor 546 | 413, 522, 543, 559 |
| poly(A) tail see mRNA | aspartyl (acidic) proteases 42 |
| polyadenylate polymerase 146 | carboxypeptidase 404 |
| polyamines 151, 433 | oligomeric ATP-dependent proteases |
| polyketides 163–165 | 405 |
| polyomavirus 538 | ClpAP/ClpXP 405 |
| polypeptide exit channel <i>see</i> ribosome | FtsH 243, 245, 405, 410–411 |
| polyproline helix 31, 459–460, 530 | HslUV (ClpQ) 405, 407-410 |
| porins 96, 427–428 | Hsp100 243, 391–393, 405, 407–409 |
| OmpG 427–428 | LonA/B 405 |
| α-hemolysin 428 | proteasome see proteasome |
| posttranslational modifications 323 | unfolding 403–406 |
| positive-inside rule 89–90, 95, 443, 566, | Ntn-hydrolases 416–417 |
| 568 | papain 404, 523 |
| potassium channels see ion channels | pepsin 8, 404 |
| PPIases (proline <i>cis-trans</i> isomerases) 386–387, | serine proteases 50–51, 229, 522 |
| 395, 402 | catalytic triad 50–51 |
| cyclophilin (CypA) 386–388 | chymotrypsin 3, 50–51, 404, 417, |
| FKBP 386–387, 395 | 566 |
| parvulin 386–387 | subtilisin 49–51, 404 |
| prediction of | trypsin 229, 404, 417 |
| disordered segments 567-569 | protease inhibitors 418–421 |
| function 574 | antichymotrypsin 481, 421 |
| secondary structure 562–569 | antithrombin 224, 418 |
| tertiary structure 569–571 | antitrypsin 418–419, 421 |
| topology of transmembrane proteins | ovalbumin 418 |
| 566–567 | plasminogen activator inhibitor 418 |
| preinitiation complex see transcription | serine protease inhibitors (serpins) 224, |
| prenol lipids 164–165 | 418–421 |
| Pribnow box 308 | reactive center loop (RCL) 418-421 |
| primary transport 425, 433–434 | proteasome 26S 405, 413–416 |
| primase RNA polymerase 276, 286 | 11S activator complex 416–417 |
| prion protein 59, 61, 418 | 19S regulator complex 413–416 |

immuno-proteasome 416 proton motive force 246–247, 252, 254, 444 11S/PA26 complex 416 PSI-BLAST 569 proteolytic chamber 20S 407, 409, 413-417 pucker see sugar pucker pulmonary surfactant membrane 203 unfolding 403–406 Protein Data Bank (PDB) 25, 42, 63, 69, 555, pulsed field gradient NMR 186, 204 purine bases see nucleosides 558-559 purple membrane 71 protein denaturation 40, 118, 386, 390 protein disulfide isomerase (PDI) 386, 388-390 pyrimidine bases see nucleosides DsbA, DsbB, DsbC, DsbD 388–389 pyrrolysine 15 protein dynamics 208 protein folding see folding process quadruplex see RNA (structural motifs) protein kinases 194, 395, 452, 456–458, 462, 464 quasi-equivalence 537–538 Ser/Thr kinases 452, 457 MAP kinase (ERK) 462, 464–465 Rab see G-proteins tyrosine kinases see also receptor tyrosine Rad51 recombination proteins 244, 301–302 kinases (RTKs) 455, 457-464 radicals 233, 295 cyclin-dependent kinase 457, 463 radical generator see ribonucleotide reductase JAK kinases 453, 455–456, 462 Raf (MapKKK) 464–465 src kinase 458–460, 533 Ramachandran plot 23–25, 27, 31, 555, 557 kinase C 194, 458, 460 Ran see G-proteins Ras see G-proteins protein kinase domain 455–456, 459–464 protein misfolding disorders (PMDs) 59, 61 RasGAP see GAP reaction center see photosynthetic reaction protein modifications acetylation 312–315 glycosylation 90, 218–220 RecA 243–244, 276, 278–280, 301–305 receptor tyrosine kinases (RTKs) 453, 455, hydroxylation 18 methylation 312-313, 315 461-467 phosphorylation 312, 391, 435, 451–452, epidermal growth factor (EGF) receptor 457, 459–461, 463, 527 462-464 SUMOylation 312, 314 insulin receptor 458, 461-462, 464 ubiquitination 312, 314 recombination 244, 272, 275, 297, 299, 301–305 oxidation 388–390 RecQ helicase see helicases protein phosphatases 452, 457–458 regulatory light chain (RLC) see myosin protein secondary structure 28-35, 562-569 replication (DNA synthesis) 3, 5–7, 105, 113, propensity 29, 562-563 118, 243, 269, 272–301, 535 protein synthesis see translation accuracy 287-288 protein tertiary structure 35, 37, 42–51, Cdc6 273, 275, 280-281 569-571 clamp loader 243, 273, 285-287 protein turnover 404 CMG complex 280–281 protechuate 3,4-dioxygenase 53 direction of synthesis 274-276, 280-281 proteoglycans 214, 218, 221, 223–224, 507 lagging strand 275, 280-281, 286-288, 294, proteolysis 219, 403–404, 418–419, 507 proteolytic chamber see proteasome leading strand 274, 280–281, 286, 288, protofilament 502 proton gradient 246-249, 252, 438-440 MCM complex 273, 276, 280-281

| Okazaki fragments 275, 286, 298 | 23S rRNA 132, 136, 139, 141, 144–145, |
|---|--|
| origin of replication 243, 275–276, 279–281, | 361–362, 365, 373–377, 379 |
| 286, 298 | 28S rRNA 361, 362 |
| origin recognition complex (ORC) 273, | 5S rRNA 134, 136–137, 329, 361–362, |
| 275, 280–281 | 365, 367 |
| primer 273, 276, 286–287, 289–294, | 5.8S rRNA 361–362 |
| 299 | A1493, A1492, G530 372–373 |
| replication bubble 282 | sarcin-ricin loop (SRL) 133, 136, 147, 373 |
| replication fork 273–277 | A2662 373–374 |
| replication complex (replisome) 276, 278, | ribosome |
| 280–281, 286 | 30S subunit 141, 361–363, 368 |
| reverse transcription see DNA polymerase | 40S subunit 361–362, 366 |
| sliding clamp 273, 276, 285–287, 295 | 50S subunit 141, 143–144, 159, 361–363, |
| repressors see transcription | 366 |
| reproduction 1–2 | 55S ribosomes (mammalian mitochondria) |
| resolvase 561 | 362, 367 |
| respiratory chain 246 | 60S subunit 361–362, 366 |
| retina 474 | 70S ribosomes (archaeal and bacterial) |
| retinal 438, 472–474 | 362–363, 382 |
| all <i>trans</i> 472–474 | 80S ribosomes (eukaryotic) 362, 366 |
| 11– <i>cis</i> 472–474 | A-site 344–345, 363, 368, 372, 374–379, |
| reverse transcriptase (RT) see DNA | 381 |
| polymerase | decoding site 363, 374, 380–381 |
| RF1, RF2, RF3 release factors see translation | E-site 368–369, 376–377, 382 |
| rhinovirus 546 | hybrid sites 364, 376, 378 |
| Rho see G-proteins | polypeptide exit channel 375–376, 402 |
| rhodopsin see G protein-coupled receptors | peptidyl transfer site 372, 379 |
| ribonucleases (RNases) 141, 146, 152, 277 | P-site 362, 367–369, 375–379, 382 |
| MRP 141 | ribosome-nascent chain complex (RNC) 92-93 |
| RNase P 152 | 102 |
| ribonucleotide reductase (RNR) 233-240 | ribosome recycling factor (RRF) 381-382 |
| active site 236–240 | ribozymes 122, 138–139, 152–159, 347, 375 |
| allosteric regulation 234, 236 | hammerhead ribozyme 138-139, 153-157 |
| overall activity site 237 | P4-P6 139, 156–159 |
| radical generator 235 | ricin 49, 147 |
| RNR classes 233, 235–237 | RNA (ribonucleic acid) |
| specificity site 233, 236–237, 239 | A form 122, 126, 135 |
| thiyl radical 233-234, 240-241 | deep groove (major groove) 122-123, 127, |
| ribosomal proteins 47, 136, 141, 147, 310, 351, | 151, 157 |
| 361, 365–367, 375, 378, 402 | double helix 121-122, 130-133, 137-144, |
| ribosomal RNA 122, 128, 130, 136, 138, 143, | 276, 558 |
| 147, 152, 307, 329, 351, 361, 366–367 | hairpin structure 122, 128–129, 132–134, |
| 16S rRNA 121, 131, 134, 139, 361–362, | 139, 152 |
| 366–368, 373 | internal loops 135–137 |
| 18S rRNA 361, 366 | junctions 138–140 |

| modified bases 128–130 | RNAse see ribonuclases |
|--|--|
| secondary structure 123, 130-132, 136-137, | RNA world 152, 233 |
| 141–143, 146–148, 153–154, 157–158 | RNR see ribonucleotide reductase |
| shallow groove (minor groove) 122, 135, | rod-shaped virus particles 537 |
| 139, 141, 144–145, 156–157 | Rossmann fold 46, 257, 264, 354, 411, 503, 512 |
| single stranded 111, 120-122, 130, 133, 148, | rotamers 25–27 |
| 547 | rotavirus 536 |
| structural motifs | Rous sarcoma virus 460, 561 |
| A-minor motif 143–145 | rRNA see ribosomal RNA |
| bulge 138–139 | rudder see RNA polymerase |
| G-tetrads 127–128 | RuvB see helicases |
| K-turn or kink-turn 141 | |
| pentaloops 132–134, 142 | saccharolipids 163–165 |
| pseudoknot 127, 142–143 | S-adenosyl methionine 63, 235, 313 |
| quadruplex (tetraplex) 127–129 | sarcin 147 |
| tetraloops 132–134, 156–159 | sarcin-ricin loop see ribosome |
| ANYA 134 | SAXS 169, 187 |
| GAAA 134, 156–159 | SCOP see fold databases |
| GNRA 133–134, 157 | SCOR database 134 |
| UNCG 134 | scoring matrices 570 |
| triplets 127–128 | secondary structure see protein secondary |
| U-turn 133–134, 150–151, 154 | structure or RNA (double helix) |
| tertiary structure 120–123, 127, 132, | secondary structure prediction 142, |
| 143–144, 148–149, 153, 159 | 562–569 |
| triple helix 127–128 | protein 562–569 |
| RNAi 307 | RNA 130–132, 141 |
| RNA polymerase 146, 219, 276, 282, 307–308, | secondary systemic amyloidosis 59 |
| 310, 325–346, 361, 479 | secondary transporters 425–426, 437, |
| active site 327–328, 331–334, 340, 342–347 | 439–449 |
| backtracking 326-327, 345-346 | secretin-like GPCR family see G protein- |
| bacterial 327–329 | coupled receptors |
| bridge helix 332–333, 343–345 | selectivity filter see ion channels (potassium |
| clamp 327, 332–333, 340, 342–343, 346 | channels) |
| DNA-RNA hybrid 326–328, 333, 343–344, | selenocysteine 15, 369 |
| 346 | senile systemic amyloidosis 59 |
| eukaryotic 326, 329–346 | sequence alignment 553-554, 564, 569-575 |
| pol I 329 | multiple (RNA) 130, 147 |
| pol II 329–346 | protein 563–564, 567 |
| pol III 329 | sequence families 560 |
| fork loops 333, 343 | sequence patterns 564, 572–573 |
| hybrid helix 326, 343, 345–346 | serine proteases see proteases |
| jaws 331–332, 342–343 | serine protease inhibitors see protease |
| RNA dependent 307 | inhibitors |
| trigger loop/helix 333, 344–346 | Ser/Thr kinases see protein kinases |
| wall 333, 340, 343 | serum amyloid A 59 |

| seven-transmembrane helix proteins see | Stahl F 272 |
|--|---|
| G protein-coupled receptors <i>or</i> | Stanley WM 8 |
| bacteriorhodopsin | stacking 20, 59, 115, 121–124, 132–135, 141, |
| SH2 domains 456, 458–463, 465 | 151, 154–157, 208, 239, 358–359 |
| SH3 domains 199, 458–461, 465 | STAT proteins (signal transducers and |
| Shc adaptor protein 458 | activators of transcription) 453, 456, |
| Shine-Dalgarno region 368 | 462 |
| sHsps see small heat shock proteins | stathmin 503 |
| σ factor see transcription factors (general) | steroid hormones 322 |
| signal recognition particle (SRP) 92–93, 139, | sterol lipids 163–165 |
| 401 | cholesterol (CHOL) 3, 70, 163, 172-173, |
| signal transduction 201, 257–258, 425, | 186–188, 201–208 |
| 451–479 | "bad" 208 |
| signaling pathways 63, 257, 451–453, 456, 459, | "good" 208 |
| 461–463, 467, 469–470, 473, 478 | stop codon see mRNA |
| immediate effects see GPCRs | strand-exchange 244, 301 |
| lasting changes see protein kinases | stratum corneum 175–176 |
| silk 42 | streptavidin 253 |
| Singer SJ 199, 201 | streptolydigin 344–345 |
| siRNA 122 | stress fibers 491 |
| sliding clamp see replication | structural alignment 553 |
| small heat shock proteins see heat shock | structural convergence 572 |
| proteins | structural genomics 574 |
| snake venoms 166 | structural superposition 553 |
| SNARE proteins 196–197 | subtilisin see proteases (serine proteases) |
| son of sevenless (Sos) 465–466 | suicide inhibitors 418 |
| southern cowpea mosaic virus 542 | sugar pucker 111, 117, 122–123, 133, 151, 558 |
| spermidine 151 | C2'-endo 111-112, 117, 123, 151 |
| spermine 151 | C3'-endo 111–112, 117, 122–123 |
| sphingolipids 163–165, 167, 197, 201, 204 | Sumner JB 8 |
| ceramides 197 | SUMOylation see protein modifications |
| glycosphingolipids 204 | superfamilies 43, 199, 242, 392, 409, 413, |
| cerebroside 193 | 439–441, 452, 559–560 |
| phosphosphingolipids 172, 182, 187, 201, | superoxide dismutase 19, 49 |
| 204 | superresolution fluorescence microscopy |
| spire 484–488 | 201 |
| FYVE domain 486, 488 | SV40 538 |
| KIND domain 486–488 | Svedberg T 8 |
| splicing 139, 152, 156, 323, 330, 346–348, 508 | SwissModel 570 |
| spliceosome 139, 346–348 | SwissProt 572–574 |
| U1, U2, U4, U5, U6 346–348 | Symmetry 8, 52–54, 65, 90–91, 107, 124, |
| sponge phase, L ₃ 181 | 126, 250–251, 254, 261–262, 278–279, |
| src kinase see protein kinases (tyrosine | 286, 316, 318–319, 338, 377, 410–411, |
| kinases) | 413, 415–416, 430, 441, 443, 447, 505, |
| S-S bond see disulfide bond | 535-540, 542, 544, 548 |

| icosahedral symmetry 53, 537–540, 544 | torsion angles 20, 23, 25–27, 33, 110–112, 114, 386, 555 |
|--|--|
| two-fold 90-91, 377, 411, 443, 505, | trans see configuration |
| 537 | transactivation domains 316, 323 |
| symporters see tranporters | transcription (RNA synthesis) 3–5, 7, 105, 110, 146, 219, 269–271, 273, 276, 282, 299, |
| T4 lysozyme see lysozyme | 307–347, 386, 412, 451, 453, 456, 464–465, |
| TAFs see transcription factors (general) | 526 |
| talin 515 | activators 307-308, 316-326, 330, 334-335, |
| Tanford C 168 | 337–338, 341 |
| Tat protein 137 | bacterial repressors 307–308, 316–325, |
| TATA-box see transcription | 452 |
| TATA box-binding proteins (TBP) see | arc 325 |
| transcription factors (general) | cro 318–320 |
| TBP-associated proteins (TAFs) see | lambda 318–319, 321 |
| transcription factors (general) | trp 317, 320 |
| T-cell receptors (TCR) 521, 528, | hybrid helix 326, 343, 345–346 |
| 530–533 | preinitiation complex 330, 333–336 |
| CDRs 522, 524, 527, 530 | recognition element 340 |
| telomerase 288, 298–301 | start site (TSS) 301, 307–308, 312, 333, 337, |
| telomerase RNA (TER) 299–300 | 339 |
| telomerase protein (TERT) 299–300 | TATA box 308, 334–335, 337–338, 340–341, |
| telomeres 128, 298–299 | 343 |
| ternary phase diagram 172–173, 203–204 | transcription bubble 326–327, 329, |
| TF (trigger factor) see chaperones | 333–335, 337, 340, 342–343 |
| TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, TFIIS | transcription factors (general) 335–341 |
| see transcription factors (general) | Mediator 341–342 |
| thermal denaturation 40, 118 | σ 329, 330, 335, 336 |
| thermophilic bacteria 40, 166 | TFIIA 330, 335–336 |
| thermosome 396 | TFIIB 330, 333, 337, 339–341 |
| thioredoxin 235, 388, 390 | TFIID 323, 330, 334–340 |
| threading methods 569 | TAFs 334–335, 337–338 |
| thrombin 224 | TAF1 338 |
| thylakoid membranes 90, 162, 194 | TBP (TATA box-binding protein) 330, |
| thymine 105, 107–109, 112–115, 121–122, | 334–335, 337–342 |
| 307 | TFIIE 330, 335–336, 340–341 |
| thymidine kinase <i>see</i> deoxyribonucleoside | TFIIF 329–330, 334–337, 339 |
| kinases | TFIIH 276, 323, 329, 335–336, |
| tymosin β4 48, 484–486 | 340–341 |
| TIM barrel 45–46, 48–49, 222, 265 | transcription factors (gene-specific) 308, 312, |
| titin 494 | 316–325 |
| $T_{\rm M}$ main transition temperature 170, 187, 203, | basic-leucine zipper (bZIP) proteins |
| 205 | 317 |
| tobacco mosaic virus (TMV) 53, 537 | GCN4 317–318 |
| topology diagrams 89–91 | MyoD 317–319 |
| T | , |

| Fos 317–318 | IF1 368 |
|---|--|
| GAL4 317 | IF2 368, 378 |
| HNF-3 | IF3 368, 382 |
| homeodomains 317, 320–321 | nascent polypeptide 90, 93, 365, |
| MAT a1 320–321 | 375–376, 401–402 |
| MAT α2 320–321 | peptidyl transfer 361–363, 365, 368, |
| Oct-1 321 | 372, 375–377, 379 |
| Jun 317–318 | proofreading 374 |
| MAX 317, 319 | release/termination factors |
| Myc 317–319, 325 | 379–381 |
| p53, p63, p73 323–324 | RF1, RF2, RF3 379-381 |
| pioneer 316–317 | ribosome recycling factor (RRF) |
| FoxA 316 | 381–382 |
| PU.1 316-317 | switch I and switch II 370-371, 374, |
| zinc fingers 321–322 | 378–379 |
| glucocorticoid receptor 322–323 | tetracycline resistance (TetM etc) |
| Zif268 317 | 369 |
| transdermal delivery 175 | translocation 346, 362, 364, 368–369, |
| transducin 468, 470–475 | 377–379 |
| transient interactions 9 | translocon 93–95, 101–102 |
| transition state 155–156, 228, 232–233, 244, 259, | transmembrane gradient 247 |
| 293–294, 467, 472–473 | transmissible spongiform encephalopathy |
| transition state analogs 231–232, 472 | (TSE) 61 |
| translation (protein synthesis) 351–382 | transporters |
| codon-anticodon interaction 368, | channel proteins 55, 81, 426, |
| 372–373 | 427–433 |
| A1492, A1493, G530 372–373 | gated or ionotropic receptors 427, |
| elongation factors 136, 147, 152, 242, | 430–433 |
| 362, 368–379 | metabotropic receptors (e g |
| EF-G, EF2 136, 242–243, 368–369, | GPCR) 427, 467–479 |
| 377–379, 382 | primary 425, 433–439 |
| EF-Ts, EF1B 369, 374–375, 378 | pumps 425–426, 429, 433–435, |
| EF-Tu, EF1A 136, 152, 243, | 438–439 |
| 362–363, 368–372, 374–375, | P-type ATPases 434–436 |
| 378–379, 381 | Ca ²⁺ -ATPases 434–436 |
| EF-Tu-EF-Ts complex 374 | H^{+}/K^{+} -ATPases 434–436 |
| LepA, EF4 369 | Na ⁺ , K ⁺ -ATPase 434–436 |
| fidelity 353, 373 | ABC transporters 436–438 |
| initial tRNA binding/selection 362, | secondary 425, 439–449 |
| 364, 368, 371–372 | antiporters 440–446 |
| initiation factors | NhaA 440–443 |
| eIF2 368 | small multidrug resistance |
| eIF4A 146 | (SMR; EmrE) 442–444 |
| eIF4B 146 | resistance-nodulation-division |
| eIF5B 368 | (RND; AcrB) 444–446 |

symporters 446–449

| neurotransmitter sodium | • |
|--|---|
| symporters (NSS; LeuT) | ubiquitin 49, 314, 405, 411–418 |
| 446–449 | iso-peptide linkages 412 |
| transthyretin (prealbumin) 53, 59, 61 | ubiquitin-activating protein (E1) 411–412 |
| triangulation number 538–540 | TAF1 338, 340 |
| tricyclic antidepressants 446, 449 | ubiquitination 312, 314 |
| trigger factor (TF) see chaperones | ubiquitin-conjugating enzyme (E2) 411–412 |
| triose phosphate isomerase (TIM) 46, 49 | ubiquitin-protein ligase (E3) 411–412 |
| tRNA 147–152 | ubiquinone 161 |
| acceptor arm/stem 148-149, 151, 352, | ultracentrifuge 8, 361 |
| 356, 358, 360, 382 | unsaturation (of lipid acyl chains) 180, 186, |
| aminoacylation 353, 355, 360 | 188, 191–192 |
| anticodon 147–148, 150–151, 352, 355, | Unwin N 8 |
| 358–360, 368, 371–373 | uracil 87, 112–113, 121–122, 126, 129–130, 296, |
| anticodon stem and loop (ASL) 352, | 300, 307, 359, 548 |
| 359, 371–372 | urease 65 |
| CCA end 358, 371, 376 | Urey H 6 |
| cloverleaf 148, 352 | UTR (untranslated region) see mRNA |
| D-stem and loop 148–151, 352, 372 | U-turn see RNA (structural motifs) |
| initiator (fMet) tRNA 362, 368 | UvrB 276 |
| L-shape148–151 | |
| mimicry 378, 381–382 | vacuoles 3 |
| modified bases 128, 130, 148 | validation 24, 555 |
| synthetases, see aminoacyl-tRNA | van der Waals' interactions 38, 168 |
| synthetases | variola (smallpox) 536 |
| tRNA-mRNA interaction 351–352, | VAST 560 |
| 361–362, 367–373, 377, 381 | VCAM-1 512 |
| cognate 356 | vesicle vesicle-mediated transport 194 |
| non-cognate 356 | vesicular stomatitis virus 546 |
| wobble base pairing 125–126, 352, | viruses |
| 373 | assembly 538-542, 547-548 |
| T-stem and loop 149–151, 352 | coat proteins 47–49, 54, 536–537, |
| V (variable) loop 148–151, 352 | 540–541, 546 |
| tropocollagen 507–508 | enveloped viruses 535, 537 542, |
| tropomyosin 55, 58, 491, 494–495, 499, | 544 |
| 501 | entry mechanisms 546–547 |
| muscle 494 | virus-cell fusion 194 |
| non-muscle 494 | visual system 452, 472–477 |
| troponin C, I, T 494–495, 497 | rod cells 472–473, 475 |
| trypsin see proteases (serine proteases) | vitamins 161, 436–437 |
| tubulin 54, 502–505 | B12 |
| tumor necrosis factor 48–49 | K 161 |
| turnip yellow mosaic virus (TYMV) 142 | vitronectin 512 |
| type II diabetes 59 | voltage-gated channels see ion channels |

tyrosine kinase see protein kinases

Walker A motif 242
Walker B motif 243
WASP (Wiskott-Aldridge syndrome protein)
485–486, 490–491
water molecules 20, 37, 39, 41–42, 53, 61–62,
70, 123, 135, 143, 148, 151–152, 181, 231–232,
240, 253, 259, 264, 344, 362, 371, 373–374,
417, 429–431, 435, 448, 467, 472, 495, 514–518

bound 41 internal 41–42

Wilkins M 7, 107–108

Watson JD 7, 105–109 Watson-Crick base pairing *see* base pair WH2 (WASP homology 2) domain 485–486, 488–489, 491 white blood cells 522, 524, 527 Willstätter R 8 winged helix-turn-helix motif 321 Woese C 152, 351

xenon 39, 66 X-ray crystallography 109, 140, 444, 555, 568 X-ray diffraction 69, 106, 202 X-ray scattering 169, 189, 208

Z-disc see muscle Z-DNA 109, 116 zeolite-like 169 zidovudine Zif268 see transcription factors (gene-specific) zinc finger 321–323, 325, 394, 486 Z-score 560–561